MOUNTAIN PEAKS

How climate change is increasing plant species richness at high altitudes **PAGE 231**

ASTROPHYSICS

PROBING THE UNIVERSE

The next big questions for gravitational waves

PAGE 164

TECHNOLOGY

DRIVER AWARENESS

Autonomous vehicles need to keep a human touch

PAGE 169

EVOLUTION

GENERATION GAME

An evolutionary history of vertebrate RNA viruses

PAGES 182 & 197
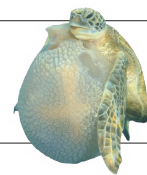
Watch the ocean

Long-term monitoring is essential for working out how changes in the Atlantic Ocean current system will affect the planet.

The Atlantic meridional overturning circulation (AMOC) has spurred scientific interest and human imagination for decades. A complex and fundamental system of ocean currents, including the wind-driven Gulf Stream, the AMOC influences the exchange of heat between the tropics and high latitudes. Driven mainly by cold, dense water in the salty Greenland and Labrador seas sinking to the bottom of the North Atlantic Ocean, the circulation regulates temperature and so serves as a global thermostat.

But for how much longer? Potential sharp changes in the circulation have been identified as a possible tipping point in Earth's physical systems. Since the 1950s, geologists and oceanographers have been gathering convincing evidence that alterations in ocean circulation are a key determinant of climate change.

Ice-core records from Greenland suggest that abrupt shifts in circulation strength triggered dramatic temperature fluctuations during the last glacial period. Climate fluctuations on such a scale have, fortunately, not occurred in the present Holocene interglacial era. Still, signs of a markedly weakening AMOC, reported in 2005 (H. L. Bryden *et al. Nature* **438**, 655–657; 2005), provoked concern that the circulation might be on the brink of tipping into a weak phase once again, possibly as a result of human-induced climate warming.

Subsequent ocean observations, from arrays of sensors strung across the North Atlantic, offered a more reassuring picture: the current was hugely variable, and so a single snapshot could be unrepresentative.

Researchers have now gone back and taken another look. In a paper in *Nature* this week, scientists present palaeo-oceanographic evidence that deep convection of surface waters in the North Atlantic — the engine that keeps the AMOC in constant motion — began to decline as early as around 1850, probably owing to increased freshwater influx from Arctic ice that had melted at the end of a relatively cold period called the Little Ice Age (D. J. R. Thornalley *et al. Nature* **556**, 227–230; 2018). This could have caused a weakening in the ocean circulation.

In a second paper, researchers used global climate models and data sets of sea surface temperature to date the onset of the weakening to more recent times, around the mid-twentieth century (L. Caesar *et al. Nature* **556**, 191–196; 2018). According to their models, the slowdown was about 15%; was most pronounced during winter and spring; and has led to a cooling of sea surface temperatures in parts of the northern Atlantic, together with a slight northward shift of the mean Gulf Stream path. This, the authors say, is probably a consequence of anthropogenic climate change.

Importantly, the findings agree that the AMOC is in a relatively weak state. The wide margin of disagreement between the two independent studies on when the circulation started to weaken is probably due to the different methods used — and it highlights how immensely difficult it is to capture the AMOC's past variability. This will probably frustrate those who prefer their science to send a clear signal. But then, science is rarely so obliging. Can the effects of climate change and natural variability on the AMOC be disentangled? And if the ocean circulation is sensitive to climate change, as is highly likely, will the currents respond abruptly and perhaps violently at some point, or will the transition be smooth? These are among the most pressing questions in climate science.

The slow progress on answering them should offer a stark reminder that the oceans are the most under-sampled component of the Earth system. The AMOC is just one part of a world-spanning circulation system, the physics — and influence on chemical cycling — of which is only poorly understood.

Numerical models are an indispensable tool for studying ocean circulation and climate. But despite ever-increasing computer power, models fall short when it comes to reconstructing something as nuanced and variable as ocean circulation. Long-term, serial measurements of circulation strength are what is needed.

It is crucial, therefore, that existing ocean monitoring systems — including the Overturning in the Subpolar North Atlantic Program and the South Atlantic Meridional Overturning Circulation programme — are maintained over decades to come. Data from these arrays of monitoring instruments are just beginning to shed light on the complex water flows in key ocean regions. Yet securing funding for lengthy studies is an ongoing fight.

There is more to be done. A United Nations sustainable development goal already includes a call for greater research capacity for promoting ocean health. Regional and national ocean-observation efforts should be coordinated, ideally under the Global Ocean Observing System. Meticulous observation is a prerequisite for understanding the oceans on which, ultimately, humankind depends. ■ **SEE NEWS & VIEWS P.180**

Cosmic sirens

Gravitational waves could help us understand differing measurements of the Universe.

Cosmology has come a long way since Edwin Hubble determined the rate of cosmic expansion around 90 years ago. Since the 1990s, multiple independent techniques have converged on values much lower than Hubble's. They differ by less than 10%, but the differences seem to be statistically significant (3.7 standard deviations). Innovative techniques, including the detection of gravitational waves from stellar collisions such as one that astronomers witnessed last August, should settle the question in the next few years. The answer could contain some new and unexpected physics.

In our expanding Universe, a galaxy's rate of recession from our own can be measured easily from its redshift — how much its light waves stretch as they travel, owing to the expansion of the intervening space. The difficult part is measuring the galaxy's distance. With his early

techniques, Hubble discovered that most galaxies seem to recede at a rate proportional to their distance. His 'Hubble constant' quantifies that proportion. Today's state-of-the-art observations suggest that, on average, galaxies' speeds increase by 73.5 kilometres per second for every megaparsec (3.26 million light years) of distance. Thus, for example, galaxies 100 megaparsecs away recede at around 7,350 km s^{-1}.

This value of the Hubble constant comes from observing stars that act as standard candles. These have known intrinsic brightness, so their distance can be estimated from how bright they look in the sky. But the value of 73.5 clashes with the 66.9 estimated in 2015 by cosmologists who mapped the cosmic microwave background — the relic radiation from the Big Bang — using the Planck observatory of the European Space Agency (ESA). The discrepancy could still turn out to be caused by unknown artefacts of the measuring techniques, but both camps say that they are increasingly confident in their results.

The Planck estimate relies on what is known as the standard model of cosmology. It makes assumptions regarding the composition of the Universe, and in particular the content of dark matter and the nature of dark energy, the mysterious driver of the acceleration of the cosmic expansion. So, if the discrepancy holds up, it could point to entirely new physics, implying that dark matter is stranger than physicists had assumed, or that the effects of dark energy change with time.

By contrast, some wonder whether standard candles might not be as reliable as astronomers think. This month, another ESA mission, the Gaia telescope, will release a 3D map of the Milky Way that has unprecedented precision and depth, and will help astronomers test the reliability of these cosmic signposts. But, ideally, astronomers would like to have more direct ways of measuring distances outside our Galaxy.

Enter gravitational waves. These stand ready to address some classic astronomical challenges with strong new evidence, as described in a News Feature on page 164. They might also help to resolve the issues surrounding the cosmic expansion. Health warning: these possibilities are speculative and controversial.

When two cosmic orbs — such as the neutron stars seen merging last August — spiral into each other, they emit gravitational waves that carry information about their distance, constituting a 'standard siren'. This enabled physicists at the US-based Laser Interferometer Gravitational-Wave Observatory (LIGO) to calculate the Hubble constant.

"If the discrepancy holds up, it could point to entirely new physics."

They obtained a value of 70, smack in the middle of the standard-candle and cosmic-microwave-background estimates. LIGO's data point has a large margin of error, but, as researchers collect more of these events, the results might end up leaning conclusively one way or the other.

Ultimately, gravitational waves could enable researchers to measure not just the current cosmic expansion, but also how the rate of expansion has evolved over the aeons. Two upcoming ESA projects will help enormously, especially if they get to fly at the same time, as many researchers hope. The gravitational-wave detector LISA (Laser Interferometer Space Antenna) should detect mergers of black holes across the Universe's history. And some astronomers anticipate that the X-ray observatory Athena (Advanced Telescope for High-Energy Astrophysics) might pick up photons from the same events and help researchers find the corresponding galaxies' redshifts — although others consider this a long shot.

Mapping standard sirens in this way should shed light on the nature of dark energy — cosmologists' most coveted goal. They hope that it will provide hints about the future of the Universe. Predictions for an infinitely long-lasting future are outside the realm of science. But cosmologists could still work out whether cosmic expansion will continue to accelerate for the foreseeable future, or whether that acceleration might increase, stop or perhaps reverse. ∎

Awards to celebrate women in science

Female scientists are under-represented in global research. *Nature* has long argued the need for initiatives to increase their opportunities and participation — so we are delighted to announce an awards programme that aims to do both.

The two annual awards will recognize inspirational early-career female researchers and those who have worked to champion young women's and girls' participation in science. By rewarding and celebrating these achievements, we hope the programmes will contribute to a positive shift towards the equity sorely needed in the research community.

The first is called the Inspiring Science Award and will honour female scientists who have completed their PhD within the past ten years and have made an exceptional contribution to scientific discovery, as reflected in publications, poster and conference presentations, leadership, tutoring and mentoring. Candidates can be nominated by anyone in their research institute, and we encourage nominations from around the globe and across all subject areas. Our independent judging process will ensure that those working under adverse circumstances or in regions where there is limited access to scientific literature will not be unfairly disadvantaged.

The second prize, the Innovating Science Award, recognizes individuals or organizations that have led a grass-roots initiative to support increased access to, or interest in, science, technology, engineering and mathematics (STEM) for girls and young women around the globe. This backs our belief that supporting early interest in STEM worldwide is a crucial step towards sustainably increasing the representation of women in these subjects. Candidates for this award can nominate themselves.

Nominations opened on 9 April and will close on 11 June 2018. A longlist of ten nominees for each award will be announced on 24 July, and a shortlist of five will be announced on 4 September. Both awards are run by Nature Research in partnership with The Estée Lauder Companies. (Full details of the criteria and nomination processes are available at nature.com/researchawards.)

The winners of the awards will be announced in October. They will receive grants of US$10,000 to build on their efforts, and an invitation to an award ceremony. The Inspiring Science Award winner will also receive a grant of up to $5,200 to support open-access publication of their research, and the Innovating Science Award winner will receive up to $5,200 to support an event that showcases their initiative. These awards complement the existing Nature Awards for Mentoring in Science and the John Maddox Prize for promoting sound science and evidence on a matter of public interest.

Nature strives to champion and showcase the achievements of researchers, and we have a responsibility to drive positive change in the research community. Our journals are committed to supporting gender equity (see go.nature.com/2glxtdj for a collection of related content). We recognize that a huge amount must be done to overcome the many barriers that women face to entry and progression in research; these awards are just one small contribution. We look forward to identifying outstanding individuals who are deserving of these awards, celebrating their achievements and sharing their stories. ∎

Demand cancer drugs that truly help patients

Drug regulators and trial designs should assess benefits that actually matter to people with cancer, says **Ajay Aggarwal**.

Already this year, the US Food and Drug Administration (FDA) has approved or extended the use of several cancer drugs that have yet to show they will prolong life or improve its quality. Unfortunately, there is no guarantee that such benefits will be demonstrated over time, and these drugs, like most cancer treatments, increase the risk of side effects such as diarrhoea and susceptibility to infection.

In my view, regulators should ensure that drugs benefit patients before allowing them to persist on the market.

As part of my work as an oncologist, patients sometimes show me headlines that describe new cancer drugs with words such as 'game changer' and 'breakthrough'. Like my patients, I'm excited to see what therapies are on the horizon. Unfortunately, these words are rarely the ones that come to mind when I appraise evidence from clinical trials. Many trials aimed at getting drugs to market depend on surrogate end points such as slowed tumour growth. However, a drug that shrinks tumours might not help to extend people's lives. This is why most oncology drugs enter the market without clear evidence that they improve either the quality or the length of life.

In 2017, my colleagues and I completed a study of all 48 cancer drugs approved by the European Medicines Agency between 2009 and 2013 (C. Davis *et al. Br. Med. J.* **359,** j4530; 2017). Of the 68 clinical indications for these drugs (reasons to use a particular drug on a patient), only 24 (35%) demonstrated evidence of a survival benefit at the time of approval. Even fewer provided evidence of an improved quality of life for symptoms such as pain, tiredness and loss of appetite (7 trials; 10%). Most indications (36 of 68) still lacked such evidence three or more years after approval. Other groups in other regions have observed similar trends. For example, a 2015 study demonstrated that only a small proportion of cancer drugs approved by the FDA improved survival or quality of life (C. Kim and V. Prasad *JAMA Intern. Med.* **175,** 1992–1994; 2015).

Once the medicines appear on the market, companies and patient advocates argue that any delay in governments covering costs for these drugs will bring about pain, suffering and unnecessary deaths, even when benefits have not been demonstrated.

If a drug does offer benefits, clinical trials are usually the best setting for these to shine through. People with cancer who are enrolled in clinical trials tend to be younger and much fitter than the general patient population. Because side effects are often worse for older or less-fit patients, benefits might not be realized or noticed in typical care settings.

Clinical trials, drug regulation and the field of medicine are all complicated. Societal values vary by country; improved survival rates might be assessed differently for different cancers, depending on how long people diagnosed with cancer are expected to live. Studies show that people's expectations about a drug's ability to extend life often far exceed what is observed. Clinicians should have honest conversations with patients to learn what constitutes a meaningful benefit for each individual.

When we choose treatment options for advanced cancer (the main indication for new cancer-drug approvals), we must consider that toxicities related to treatments may shorten life expectancy rather than extend it, and we should ensure that treatments do not diminish quality of life. Unfortunately, many clinical studies either neglect quality-of-life measures entirely or rely on unvalidated instruments. One seminal study demonstrated that people with advanced lung cancer who had early access to palliative care alongside standard treatments had greater improvements in quality of life and survival, despite receiving fewer aggressive end-of-life treatments (J. S. Temel *et al. N. Engl. J. Med.* **363,** 733–742; 2010).

Regulators also need to focus on more measures that people value: reduced toxicity, and the ability to maintain enough function to return to work or keep up social ties.

Some argue that the time required for randomized, controlled trials with meaningful measures would take too long. However, there have been innovations in designing robust trials measuring overall survival and quality of life, even in slowly progressing diseases such as prostate cancer.

Approvals that let drugs stay in the marketplace on the basis only of quick, easy surrogate end-points are unlikely to produce highly effective treatments; we will simply get more drugs providing marginal value.

I believe that the low bar also undermines innovation and wastes money. Copycat drugs with minimal benefits will continue to be approved on the basis of surrogates, and so will minimize incentives for true breakthroughs and game changers. At the same time, a large influx of drugs bringing limited benefit will force governments to spend a greater proportion of health funding on cancer drugs rather than on other treatment options.

> ## MOST INDICATIONS *STILL* LACKED EVIDENCE OF A SURVIVAL *BENEFIT* THREE OR MORE YEARS AFTER APPROVAL.

Another risk is that emerging, heavily marketed drugs could blind clinicians and patients from looking anew at existing options that might bring bigger benefits. It amazes me how much attention is given to drugs even as people with cancer struggle to access surgery and radiotherapy. Investment in screening and diagnostics research also falls far behind that of drug research.

Ultimately, I want to access the best available therapies for the people I treat: the ones most likely to bring meaningful improvements in their quality and length of life, and the ones that reduce the toxicity associated with treatment. Any new cancer therapy, drug or not, should undergo robust evaluation for outcomes that truly matter to individuals. As it is, limited finances are too often being directed from evidence-based therapies to those that promise false hope. ■

Ajay Aggarwal *is an oncologist at Guy's and St Thomas' NHS Trust, London, UK, and a senior lecturer at King's College London.*
e-mail: ajay.aggarwal@kcl.ac.uk

SEVEN DAYS
The news in brief

AI boycott ends

A group of artificial intelligence (AI) researchers has ended a boycott of collaborations with the Korea Advanced Institute of Science and Technology (KAIST) in Daejeon, South Korea. In a petition letter announced on 4 April, more than 50 researchers stated their concerns over KAIST's Research Center for the Convergence of National Defense and Artificial Intelligence, which is operated with Hanwha Systems, a defence company in Seoul. The signatories said they would not work with KAIST until they had been assured that the centre "will not develop autonomous weapons lacking meaningful human control". On 9 April, the group rescinded the boycott after KAIST president Sung-Chul Shin said that the university would not engage in research that is counter to "human dignity including autonomous weapons lacking meaningful human control".

Scientist sentenced

A Chinese scientist was given a ten-year prison sentence by a US court on 4 April for stealing genetically altered seeds. Zhang Weiqiang, a Chinese national who is a US resident, was found guilty in February 2017 of stealing hundreds of seeds from his previous employer, Ventria Bioscience in Junction City, Kansas, and storing them in his home. Zhang was suspected of passing the seeds on to a Chinese crop-research institute. Plants grown from the modified seeds produce therapeutically valuable proteins such as human serum albumin, which is found in blood and needed in large quantities to replenish blood lost in injury or during surgery. Court documents say that Ventria Bioscience spent

HALIL EL-ABDULLAH/ANADOLU AGENCY/GETTY

Suspected chemical attack in Syria

An international team has begun an investigation into a suspected chemical-weapons attack on civilians in Douma (pictured), a besieged town in Syria. It is reported that the attack on 7 April killed dozens of people and affected hundreds more. The Organisation for the Prohibition of Chemical Weapons (OPCW), which enforces the global treaty banning the use of such arms, said on 9 April that it would send inspectors to Syria on a fact-finding mission. Inspectors will try to identify any agents used, thereby confirming whether a chemical attack occurred. Physicians for Human Rights, a humanitarian organization in New York, called for immediate independent collection of environmental and biological samples from Douma. According to Human Rights Watch, there have been 85 chemical-weapons attacks in Syria since 2013, and the OPCW has confirmed the use of mustard gas and the nerve agent sarin in some instances. Since the civil war in Syria began in 2011, an estimated 400,000 people have been killed.

millions of dollars developing the strains. Zhang was found guilty of conspiracy to steal trade secrets and conspiracy to transport stolen property across state borders, and sentenced to 121 months in prison.

Fellowship launch

UK Research and Innovation (UKRI) — Britain's powerful new research-funding body — announced the creation of a fellowship scheme for early-career scientists on 3 April. The Future Leaders Fellowships, launched in the organization's first week of operation, will by 2021 support 550 early-career researchers in any discipline, each funded for up to 7 years. UKRI says the fellowships are in addition to existing schemes run by the individual research-funding councils, which have been subsumed under UKRI. The new scheme, which is open to researchers in academia and industry, and around the world, is designed to "develop, retain, attract and sustain research and innovation talent in the UK".

Opioid research

On 4 April, the US National Institutes of Health (NIH) announced a US$1.1-billion research initiative aimed at curbing the opioid epidemic. The budget for the programme, called Helping to End Addiction Long-term (HEAL), is nearly double what the NIH spent on such research in 2016. HEAL's goals include designing longitudinal studies to follow people with chronic pain, developing neuroimaging technologies to understand how pain manifests in the brain, testing new therapies for addiction and overdoses and partnering with the US military and the Department of Veterans Affairs to develop non-pharmaceutical approaches to pain management. The NIH is also

partnering with private drug companies to develop non-addictive painkillers.

Supercomputer cash

The US Department of Energy (DOE) announced up to US$1.8 billion in funding for at least two new supercomputers on 9 April. The project request follows an award last June for the Aurora supercomputing system, currently being developed at Argonne National Laboratory in Illinois. One of the systems resulting from the April announcement will be housed at the Oak Ridge National Laboratory in Tennessee, and the second will be at the Lawrence Livermore National Laboratory in Livermore, California. The two supercomputers would come online between 2021 and 2023. The DOE says funds could also support a third system or an upgrade to Aurora, depending on need.

Defence research

The number of Japanese research institutions that have procedures to evaluate the ethics of military-related research has doubled in the past year, according to a survey released by the Science Council of Japan on 4 April. Of the 135 Japanese institutes and universities that responded to the survey, 26% said that

they now screen research for possible military applications, up from 13% in 2017. The council, which advises the Japanese cabinet, called for a boycott of military-related research in March 2017 after the government boosted funding for 'dual use' scientific research with potential military applications. In response to the funding boost, the council asked institutions to introduce evaluation guidelines for such research.

Whole genomes

The UK Biobank, which holds health records and other biological data for some 500,000 people, announced on 5 April that it plans to sequence the complete genomes of 50,000 participants. In 2017, the powerhouse database released limited genetic data — on 800,000 DNA variants that tend to differ between individuals — for all participants. These data have already been mined in dozens of studies looking for DNA variants linked to diseases and biological traits. The whole-genome sequencing, which will be funded by a £30-million (US$42-million) grant from the UK Medical Research Council, will allow for more-sophisticated studies. In January, the UK Biobank (**pictured**, a blood sample
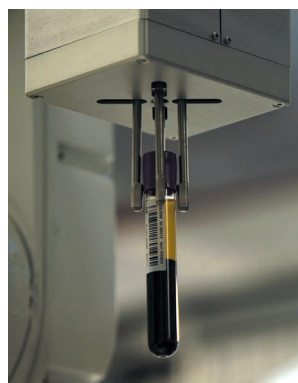
from the biobank) announced plans to release data on the exomes — the small portion of genome that encodes proteins — of all participants by 2019.

Fusion-lab site

A laboratory near Rome will host a €500-million (US$618-million) experiment on nuclear fusion called the Divertor Tokamak Test (DTT) facility. Italy's energy-research agency, ENEA, said on 4 April that it had selected its Frascati laboratory over eight other candidate sites in the country. The DTT will test technologies for extracting heat from fusion plasma. This is a necessary step in the development of fusion-based commercial power stations that the experimental ITER reactor, currently under construction in France, is not designed to take. ENEA plans to begin construction before the end of the year, and to fund the project in

part with a €250-million loan it has requested from the European Investment Bank. It also expects a €30-million contribution from China and €60 million from the European EUROfusion consortium.

Moon prize revived

XPRIZE of Culver City, California, is continuing its private competition to spur lunar exploration, even though a US$30-million version of the programme sponsored by Google ended on 31 March with no winners. The group said on 5 April that it was looking for new sponsors but would continue the competition without cash for now. Five teams were in the running for the Google Lunar XPRIZE, which required entrants to develop a vehicle capable of landing on the Moon, travelling 500 metres across the surface and broadcasting images and video back to Earth.

Article-access tool

Scholarly-services firm Clarivate Analytics has purchased artificial-intelligence company Kopernio, which has a tool that helps researchers to find and access journal articles with one click. Clarivate, which owns the scholarly search engine Web of Science, announced the deal on 10 April but did not disclose its value. Kopernio offers a browser plug-in that makes it easier for researchers to find and download literature that they already have legitimate access to. The feature will be integrated into Web of Science to help users get around the problem of having to sign in to multiple sites to access articles available as part of an institutional subscription. It also logs academics' access credentials so that they can access pay-walled papers off campus.

↻ **NATURE.COM**
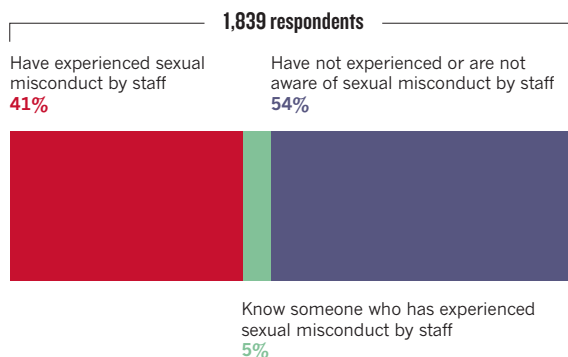For daily news updates see:
www.nature.com/news

TREND WATCH

A survey of 1,839 current and former UK students has found that 41% experienced sexual misconduct — such as inappropriate comments, unwanted touching or assault — by staff at university. About 12% of current students said a staff member had touched them in a way that made them uncomfortable. Women were more likely than men, and postgraduates more likely than undergraduates, to report such harassment. The National Union of Students conducted the survey.

SEXUAL MISCONDUCT AT UK UNIVERSITIES

In a national survey of 1,839 current and former UK students, 41% said they had experienced sexual misconduct by staff. These experiences were particularly prevalent among postgraduates.
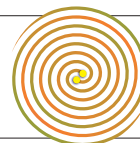
1,839 respondents

Have experienced sexual misconduct by staff
41%

Have not experienced or are not aware of sexual misconduct by staff
54%

Know someone who has experienced sexual misconduct by staff
5%

NEWS IN FOCUS

ALFRED PASIEKA/SPL
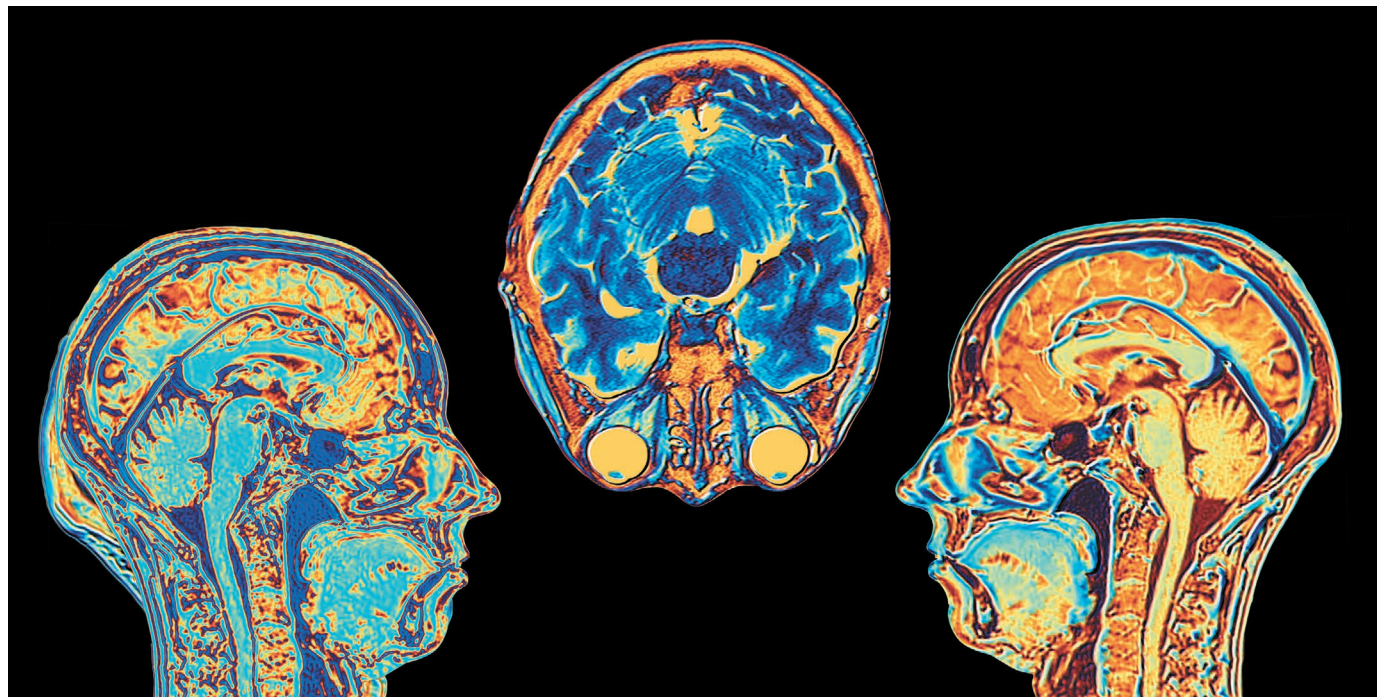
China's new brain-science centre will host some 50 principal investigators and will also support external researchers.

CHINA

Beijing launches pioneering neuroscience centre

Large research facility will be key part of much-anticipated brain initiative.

BY DAVID CYRANOSKI

For China, 2018 is shaping up to be a big year in brain science. Beijing announced plans last month to build a major neuroscience centre that will rival in size some of the world's largest organizations in that discipline. It will also serve as a core facility for the country's long-awaited brain project — China's version of the high-profile brain-science initiatives under way elsewhere in the world.

The Chinese Institute for Brain Research was officially established in Beijing on 22 March, with an agreement signed by representatives of the Beijing municipality and seven research organizations based in the capital.

The agreement named two neuroscientists — Peking University's Rao Yi and Luo Minmin of the National Institute of Biological Sciences in Beijing — as co-directors.

The new Beijing facility will be one of the first concrete developments in China's national brain-research project, which has been under discussion for five years but has yet to be formally announced. The United States and Europe each launched billion-dollar brain initiatives in 2013, and Japan set up a smaller project the following year. South Korea answered with its own initiative in 2016.

China is expected to complement these projects with its rapidly growing cadre of top neuroscientists, abundant supplies of research monkeys and big investments in brain-imaging facilities. "The brain is such a complex system that significant efforts are needed to tame this complexity at an international level," says Katrin Amunts, scientific-research director of Europe's Human Brain Project. China has the potential to provide important insights that relate to the work of other projects, she says.

PLANS AFOOT

Luo says that he will oversee the roughly 50 principal investigators who will have laboratories at the new centre, with Rao taking charge of external grants that will support around 100 investigators throughout China. Luo says that the centre will be similar in ▶

▶ organization and scientific scope to the US National Institute of Mental Health, a major US brain-science funder, although on a smaller scale.

The Chinese centre will be a partnership between Beijing's premier biomedical institutions, among them the Chinese Academy of Sciences, the Academy of Military Medical Sciences, Peking University and Tsinghua University. Luo says it will support projects that use the latest biomedical methods, such as high-throughput single-gene sequencing, precision genome editing and big-data processing. He also hopes to develop better imaging tools, including a voltage sensor that can directly record neuronal activity, and high-speed-imaging microscopes that will allow detailed views of brain activity.

This year, Luo plans to use 180 million Chinese yuan (US$29 million) provided by the Beijing municipal government to hire the first five or six research groups, and to install them in a building already constructed by the municipality, which is across the road from his institute. When operating at its full capacity of 50 researchers, which Luo plans to have within 5 years, some 400 million yuan per year will be needed. He hopes to secure this from

the brain-science project, with a substantial amount still coming from Beijing.

Luo says that it will be a "docking site" for the Chinese brain project, which has been in planning since the United States and Europe launched their programmes. So far, few firm details about the project have been released. Scientists who spoke to *Nature* say they expect that the government will officially launch the initiative some time this year.

STAFFING CHALLENGES

In the meantime, other facilities are preparing their bids for support from the national project. A large science park under construction in Shanghai will house a 'southern centre' for neuroscience research. The centre's organizers say this will support many more principal investigators than its Beijing counterpart, which scientists are dubbing the northern centre.

Feng Jianfeng, a computational biologist and head of Fudan University's Institute of Science and Technology for Brain-inspired Intelligence, has been involved in organizing the Shanghai projects. He says that one focus will use artificial intelligence (AI) to study brain diseases. Feng adds that, with 190 million yuan

from the university, he is already setting up a brain-imaging facility that will house the largest number of magnetic resonance imaging devices in Asia, and will be based at the southern centre. AI algorithms will screen the images, comparing diseased brains with healthy ones, to form part of the world's largest brain database, he says.

Another programme expected to be integral to the country's brain-science initiative is an international connectome project, which is being designed by Mu-Ming Poo, director of the Institute of Neuroscience in Shanghai. Connectome projects attempt to map out all the neural connections in the brain.

Finding enough researchers might be the greatest challenge for both the individual centres and the Chinese brain-science project. Jeffrey Erlich, a neuroscientist at NYU Shanghai, says that, as well as hiring top neuroscientists, the initiatives will need to fund postdoctoral positions and graduate-school research posts offering internationally competitive salaries.

"That would increase the number of top students going into neuroscience," says Erlich. "Then, in five to ten years, China could have a fresh crop of top young scientists." ■

Exoplanet hunter will seek worlds close to home

NASA's mission is designed to spot planets orbiting nearby bright stars.

BY ALEXANDRA WITZE

Filling the shoes of NASA's Kepler spacecraft won't be easy. Since its launch in 2009, Kepler has discovered nearly three-quarters of the 3,700-plus known exoplanets. And there are thousands more candidates waiting to be confirmed.

So NASA is taking a different approach with its next planet-hunting mission. On 16 April, the agency plans to launch the US$337-million Transiting Exoplanet Survey Satellite (TESS), which will scrutinize 200,000 nearby bright stars for signs of orbiting planets. TESS will probably find fewer worlds than Kepler did, but they are likely to be more important ones.

"It's not so much the numbers of planets that we care about, but the fact that they are orbiting nearby stars," says Sara Seager, an astrophysicist at the Massachusetts Institute of Technology (MIT) in Cambridge and deputy science director for TESS.

TESS is meant to identify planets that are close enough to Earth for astronomers to explore them in detail. Team scientists estimate that the spacecraft will discover more than 500 planets that are no more than twice the size of Earth (P. W. Sullivan *et al. Astrophys. J.* **809**, 77; 2015). These worlds will form the basis for decades of further studies, including searches for signs of life. "We'll see a whole new opening of exoplanet studies," Seager says.

"We'll see a whole new opening of exoplanet studies."

Both Kepler and TESS are designed to scan the sky for planetary transits, the slight dimming that occurs when a planet moves across the face of a star and temporarily blocks some of its glow. For most of its mission, Kepler stared at a deep but narrow slice of the Universe — peering out some 920 parsecs (3,000 light years) from Earth but covering

only 0.25% of the sky. Its celestial census showed that planets were common throughout the Milky Way. "We found that planets are everywhere," says Elisa Quintana, an astrophysicist at NASA's Goddard Space Flight Center in Greenbelt, Maryland.
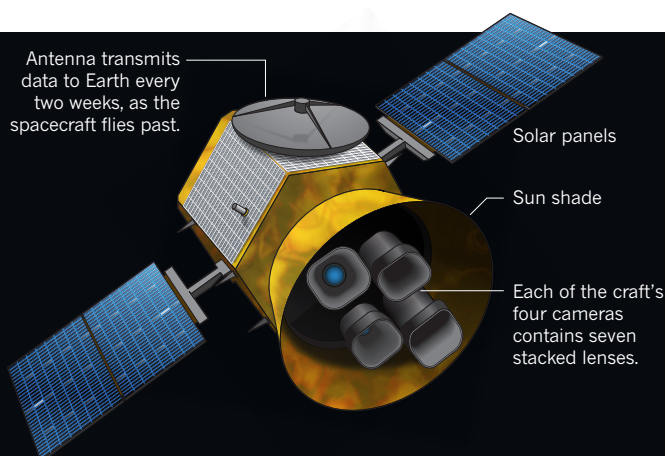
MEETING THE NEIGHBOURS

By contrast, TESS will go shallow and broad — looking at stars within 90 parsecs of Earth but covering more than 85% of the sky. Its 4 cameras will give the spacecraft a field of view about 20 times the size of Kepler's (see 'Scanning the sky'). TESS will sweep the southern sky first and then, after a year, turn its attention to northern stars.

The observing swathes in each hemisphere will overlap at the south and north ecliptic poles, which are points perpendicular to the plane of Earth's orbit. That's by design, because NASA's James Webb Space Telescope, now planned for a 2020 launch, will also be able to study those regions at any
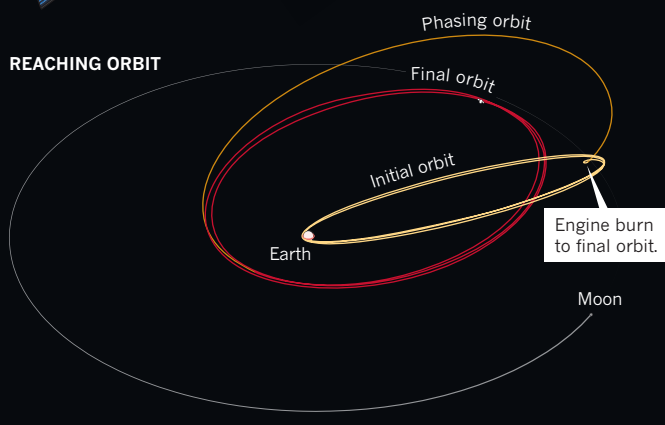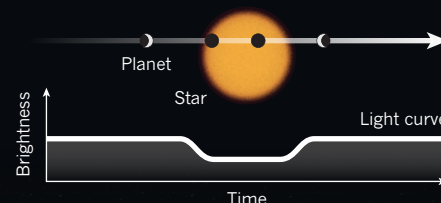
SCANNING THE SKY

NASA's Transiting Exoplanet Survey Satellite (TESS) will monitor 200,000 stars during its 2-year mission hunting worlds outside the Solar System. Researchers expect the craft to find more than 1,600 planets, including about 500 that are twice the size of Earth or smaller.

Antenna transmits data to Earth every two weeks, as the spacecraft flies past.

Solar panels

Sun shade

Each of the craft's four cameras contains seven stacked lenses.

ON THE LOOKOUT
TESS will seek new worlds by watching for the dimming that occurs when a planet passes across the face of its star.

Planet

Star

Light curve

Brightness

Time

REACHING ORBIT

Phasing orbit

Final orbit

Initial orbit

Earth

Engine burn to final orbit.

Moon

Earth

The Kepler craft searches up to 920 parsecs (3,000 light years) from Earth, but covers just 0.25% of the sky.

1,000

2,000

3,000

TESS's search area will extend 90 parsecs (300 light years) from Earth and cover 85% of the sky.

given time. Webb's 6.5-metre primary mirror will allow detailed spectroscopic studies of the planets' atmospheres, but it will be in high demand for a range of other astronomical research. "The time on Webb is going to be so precious," says George Ricker, an astrophysicist at MIT and TESS's principal investigator.

Once TESS spots interesting planetary candidates, a fleet of Earth-based observatories will kick into action to gather more data. These

will include planet-hunting stalwarts such as the HARPS instrument at the European Southern Observatory in La Silla, Chile, and the new Miniature Exoplanet Radial Velocity Array (MINERVA)-Australis, a group of five planned 0.7-metre telescopes near Toowoomba, Australia. "We have the ability to hammer on a target every night if we need to," says Rob Wittenmyer, an astronomer at the University of Southern Queensland in Toowoomba who helps lead MINERVA-Australis.

These and other ground-based telescopes will be able to deduce the TESS planets' masses, and from that their composition — whether they are rocky, icy, gassy or something else.

A WHOLE NEW WORLD
Recent research suggests that TESS may yield a greater bounty than once thought. Earlier this year, MIT astronomer Sarah Ballard recalculated how many planets TESS might find orbiting the cool, plentiful stars known as M dwarfs — and predicted some 990 such planets, 1.5 times more than earlier estimates (S. Ballard Preprint at https://arxiv.org/abs/1801.04949; 2018). The sheer volume of discoveries would allow astronomers to begin comparing broad classes of exoplanets: learning how stellar flares affect planetary atmospheres, for instance, or what sorts of planets surround stars of different ages.

TESS will soon have company. The European Space Agency (ESA) plans to launch its Characterising Exoplanet Satellite late this year. The craft will measure the sizes of known planets — from those a little bigger than Earth to ones that are roughly Neptune-sized — orbiting nearby bright stars. ESA is also planning two missions for the 2020s: PLATO to study Earth-sized exoplanets, and ARIEL to study planetary atmospheres.

The next generation of missions will come just in time: Kepler is on its last legs, with only a few months' worth of fuel left to help it make its final discoveries. ∎

The Transiting Exoplanet Survey Satellite will search more than 85% of the sky.

EQUALITY

UK wage data reveal science's gender gap

Reports affirm systemic struggles for women in science.

BY HOLLY ELSE

Many UK science employers pay women much less than men, and some institutions are far less equal than others, according to an analysis conducted by *Nature* of statistics released last week. Universities, pharmaceutical companies, funders and other science-focused organizations maintain a gender pay gap that is 50% greater than the national average for all employers.

In 2017, the United Kingdom became one of the first nations in the world to require employers to report differences in pay between men and women. Organizations that employ more than 250 people must report details of their gender pay gap, the representation of men and women in each pay quartile and the gender breakdown of who receives bonus pay. More than 10,200 organizations have now uploaded data to the government's portal for gender pay-gap figures.

The gender pay gap refers to the difference in the average hourly wage of all men and women across a workforce. It is not the same as unequal pay, when men and women are paid differently for performing the same role, which has been illegal in the United Kingdom since 1970.

To see how science shapes up, *Nature* analysed data for universities, research institutes, selected grant funders and some industrial employers (see 'Research wage gap').

Science institutions fared poorly overall. Of the 172 organizations included in the analysis, 96% pay men more than women, according to the companies' reported median pay gaps. Nationwide, 78% of all organizations favour men financially. The median gap between genders among science employers is 15%, compared with the UK median of 10%. The median offers the best representation of typical differences in pay, because it is not skewed by outlying high or low figures.

The median pay gap for universities is 16%, research institutes 9%, funders 10%, industrial employers 12% and for five scientific publishers, 22%. (Macmillan Publishers, the division of SpringerNature that publishes *Nature*, has a median gender pay gap of 13%.)

Much of an organization's pay gap comes down to how women and men are distributed through the ranks. Women are often over-represented in low-paid and low-skilled jobs, whereas men are likely to make up the bulk of workers in senior and high-paid roles. Stripping out non-academic roles from the university data would probably shrink the gap between male and female earnings, says Jeff Frank, an economist at Royal Holloway, University of London, in Egham, UK, who studies gender pay gaps in science.

The median gender pay gap for all academic staff at UK universities is 12%, according to a report released in 2017 by the University and College Union (see go.nature.com/2ezlidw). For professors at research-heavy institutions, the figure is 7%. The driving force for this pay gap, says the report, is a "very clear and continuous decline" in the proportion of women represented as academic rank increases in seniority. Across the board, 45% of the entire academic workforce is female, for example, but less than one-quarter of all professors are women.

FUNDING GATEWAY

The London-based Wellcome Trust has a 21% median gender-pay gap, which it is seeking to close by training staff to help mitigate bias, and introducing fairer ways of recruiting, promoting and retaining women at senior levels. Couch says that this approach is important because her organization and other funding agencies act as the gatekeepers to science, and the Wellcome Trust could be missing out on supporting excellent ideas. Of the other nine funders *Nature* examined, three had nonexistent or negligible pay gaps, including the Engineering and Physical Sciences Research Council and the Economic and Social Research Council.

Across 29 research companies analysed by *Nature*, oil and gas businesses, including those owned by BP and Shell, generally had the largest median pay gaps and the lowest proportions of women in the top pay quartile.

In the pharmaceutical industry, there is huge variation. One company, MSD (the UK subsidiary of Merck), has a 7% pay gap in favour of women; another, GlaxoSmithKline, reports small differences in pay, which favour men. Pfizer and AstraZeneca have more-typical gender pay gaps, at 18% and 13%, respectively.

There is one obvious way to abolish gender pay gaps. One UK university eliminated its professorial pay gap overnight — by simply boosting women's pay, says Alice Chilver, head of organizational development at University College London, who heard the anecdote in March, at a conference in London that she organized on gender pay gaps.

Quick fixes aside, universities and other science employers might need to make more-significant changes to be able to close their gender pay gaps. "It is all about the culture, the behaviour, the habits and patterns, the beliefs, the fears," Chilver says. "Until we can tap into and work with those cultures more effectively, change is going to be very slow." ■

SOURCE: HTTPS://GENDER-PAY-GAP-SERVICE.GOV.UK/

RESEARCH WAGE GAP

UK companies with more than 250 employees had until 5 April to publish statistics on the gender pay gap. *Nature* analysed data from universities, pharmaceutical companies and other employers of scientists.

1 Science employers averaged a median pay difference of 15% in favour of men, compared with the UK-wide figure of 10%.
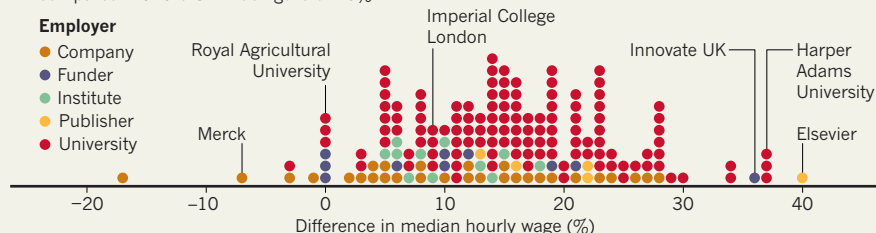
2 A lack of women in senior roles underlies many pay gaps. One report found that less than one-quarter of UK professors are female, even though women make up 45% of the academic workforce.

Gender by quartile (%) ■ Men ■ Women

	Lower		Lower middle		Upper middle		Top	
Publishers	29	71	34	66	45	55	60	40
Funders	36	64	35	65	43	57	53	47
Universities	35	65	39	61	47	53	56	44
Institutes	45	55	50	50	58	42	70	30
Companies	56	44	65	35	70	30	73	27

Nature's analysis is based on data from 122 universities (including Cambridge and Oxford colleges with more than 250 employees), 11 science institutes, 29 companies, 10 research funders and 5 science publishers.

POLICY

Puerto Rico's statistics agency in jeopardy

Reorganization could threaten reliable, independent data about the island, critics say.

BY GIORGIA GUGLIELMI

Puerto Rico's senators last week approved a plan to overhaul an independent statistics agency tasked with coordinating the collection and analysis of crucial data on the island. The reorganization will wreck the US territory's ability to produce credible data about itself, including updated estimates of the death toll from last year's Hurricane Maria, critics of the plan say.

The decision paves the way for the restructuring of several government agencies, including the Puerto Rico Institute of Statistics (PRIS). To make it official, policymakers must now approve legislation dismantling the laws that established PRIS. Under Governor Ricardo Rosselló's plan to streamline government agencies, first introduced in January, PRIS would become an office in the Department of Economic Development and Commerce, which would contract the institute's duties to private companies.

But some fear that privatizing official statistics isn't in the island's best interests. "The private companies are going to be chosen by the government and we don't know how independent their leaders are going to be," says

Changes loom for body that handles statistics such as hurricane damage.

JOE RAEDLE/GETTY

Mónica Feliú-Mójer, director of communications and science outreach at Science Puerto Rico, a non-profit group based in San Juan.

Another worry is that private companies might not distribute their data freely, or provide access to information on how they collected and analysed the numbers, says Steve Pierson, director of science policy at the American Statistical Association in Alexandria, Virginia.

Since PRIS began operating in 2007, it has worked to improve the quality of government

agencies' statistics: the institute trains statisticians in new methodologies, ensures that data collection and analysis meet international standards and helps the agencies to make their data publicly accessible.

PRIS has improved tracking of Puerto Rico's mortality rate, and it established a fraud-prevention system related to the US Medicaid health-insurance programme, saving the government millions of dollars.

But Rosselló disputes the agency's effectiveness. PRIS "has failed in establishing efficient data gathering procedures that produce reliable statistics", says Alfonso Orona, the governor's principal legal counsel. He says that outsourcing data collection and analysis will help to address this.

It's likely that lawmakers will approve the legislation that would officially dismantle the institute, says Roberto Rivera, a statistician at the University of Puerto Rico at Mayagüez. Puerto Ricans are grappling with many issues, including the aftermaths of last year's hurricanes and a series of education and labour reforms, so PRIS is not a priority, he says. "If there's not enough pressure on the government, they'll get their way." ■

THERAPEUTICS

Promising cancer drug hits snags

Physicians struggle to identify which patients are likely to respond to cutting-edge therapy.

BY HEIDI LEDFORD

Cancer specialists in the United States had high hopes last year when they gained approval for a new approach to treatment: a drug that targeted certain tumours regardless of where they first appeared in the body.

But clinicians and researchers are struggling to put that plan into practice. Although the drug itself works well against a variety of tumour types, there have been problems

with some of the tests used, which identify suitable tumours on the basis of certain molecular markers.

On 15 April at the American Association for Cancer Research annual meeting in Chicago, Illinois, researchers and representatives from the US Food and Drug Administration (FDA) will discuss how best to tackle the issue. "If you get a false negative result, you're not going to give that patient the therapy, which is terrible," says Zsofia Stadler, an oncologist at the Memorial Sloan Kettering Cancer Center in New

York City. "That's why there's such a debate."

The drug in question, pembrolizumab (Keytruda), works by firing up the body's immune responses against tumours. First approved by the FDA in 2014 to treat melanoma, it has since been given the go-ahead to treat a handful of other cancers, including lung cancer.

But last year, researchers reported that patients whose tumours had a disabled DNA-repair system also responded to the drug, regardless of where the tumour originated ▶

▶ (D. T. Le *et al. Science* **357**, 409–413; 2017). Damaged DNA can yield mutant proteins, which the immune system could target as potential invaders. Scientists think that this increases the chances that immune cells unleashed by pembrolizumab will find and attack the tumour.

In May 2017, the FDA allowed pharmaceutical giant Merck of Kenilworth, New Jersey, to market pembrolizumab to people with advanced-stage cancer who had any solid tumour with that particular DNA-repair defect. "This is absolutely a breakthrough approval," says Razelle Kurzrock, an oncologist at the University of California, San Diego. "We have seen some dramatic responses in our patients."

But the three kinds of test commonly used to look for DNA damage arising from that defect can produce conflicting results, says Heather Hampel, a genetic counsellor at the Ohio State University in Columbus. One relies on PCR, a process that amplifies specific regions of the genome; a second looks for certain proteins; and a third relies on DNA sequencing. "Which is the best? Is any positive on any test sufficient?" Hampel says. "Does that mean you should try them all? No one wants to miss a patient who might benefit from pembrolizumab."

And there are signs that some of the tests might work better in certain tissues than in others, says Shridar Ganesan, a physician and cancer researcher at the Rutgers Cancer Institute of New Jersey in New Brunswick. PCR assays, for example, look for changes in certain regions of DNA called microsatellites. Particular microsatellites might be more prone to damage in some tissues than in others, he says.

Stadler notes that the degree to which the DNA changes might also vary from tissue to tissue: colon cancers tend to accumulate many mutations, whereas tumours in the adrenocortex generally have fewer. That can lead to a false negative result in tissues with fewer mutations, she says.

> *"Which is the best? Is any positive on any test sufficient? Does that mean you should try them all?"*

Similar complications might arise for some future tissue-agnostic drug approvals, particularly those based on DNA-repair defects. This could include drugs called PARP inhibitors, which are approved in the United States for breast and ovarian cancers caused by mutations in either of two genes involved in DNA repair: *BRCA1* or *BRCA2*. Researchers are looking at whether PARP inhibitors might also work in any solid tumour that carries similar DNA-repair defects, even if they aren't caused by *BRCA1* or *BRCA2* mutations. There are multiple tests available for identifying the patterns of DNA damage in such tumours, says Hampel.

Evidence has also been building that the overall number of mutations in a tumour could indicate how likely it is to respond to immunotherapies such as pembrolizumab. Tests for this might also be complex, notes Stadler.

Eventually, some of these issues will be ironed out, says Michael Overman, an oncologist with the University of Texas MD Anderson Cancer Center in Houston, as researchers gather data on which tests work best in which cancers. But the FDA was wise to move forward with the approval rather than wait for more evidence to sort out the issues with the molecular marker tests, he says. "There are still a lot of open questions, but the therapy works exceptionally well," he says. "It was the right thing to do." ∎

CORRECTIONS & CLARIFICATIONS

The News story 'Alzheimer's study zeroes in on enigmatic protein' (*Nature* **555**, 567–568; 2018) misstated the radioactive marker that will be used in the tau scans. It is GTP1, not GPT1.

The News story 'Copyright reforms draw fire from scientists' (*Nature* **556**, 14–15; 2018) should have made it clear that when Vanessa Proudman talked of "that process" she was referring to how institutional repositories deal with copyright violations.

HERE COME THE WAVES

After a clutch of historic detections, gravitational–wave researchers have set their sights on some ambitious scientific quarry.

BY DAVIDE CASTELVECCHI

In the mid-1980s, Bernard Schutz came up with a new solution to one of astronomy's oldest problems: how to measure the distance from Earth to other objects in the cosmos. For generations, researchers have relied on an object's brightness as a rough gauge for its distance. But this approach carries endless complications. Dim, nearby stars, for example, can masquerade as bright ones that are farther away.

Schutz, a physicist at the University of Cardiff, UK, realized that gravitational waves could provide the answer. If detectors could measure these ripples in space-time, emanating from interacting pairs of distant objects, scientists would have all the information needed to calculate how strong the signal was to start with — and so how far the waves must have travelled to reach Earth. Thus, he predicted, gravitational waves could be unambiguous markers of how quickly the Universe is expanding.

His idea was elegant but impractical: nobody at the time could detect gravitational waves. But, last August, Schutz finally got the opportunity to test this concept when the reverberations of a 130-million-year-old merger between two neutron stars passed through gravitational-wave detectors on Earth. As luck would have it, the event occurred in a relatively nearby galaxy, producing a much cleaner first measure than Schutz had dreamed. With that one data point, Schutz was able to show that his technique could

become one of the most reliable for measuring distance. "It was hard to believe," Schutz says. "But there it was."

More mergers like that one could help researchers to resolve an ongoing debate over how fast the Universe currently is expanding. But cosmology is just one discipline that could make big gains through detections of gravitational waves in the coming years. With a handful of discoveries already under their belts, gravitational-wave scientists have a long list of what they expect more data to bring, including insight into the origins of the Universe's black holes; the extreme conditions inside neutron stars; a chronicle of how the Universe structured itself into galaxies; and the most-stringent tests yet of Albert Einstein's general theory of relativity. Gravitational waves might even provide a window into what happened in the first few moments after the Big Bang.

Researchers will soon start working down this list, with the help of the US-based Laser Interferometer Gravitational-Wave Observatory (LIGO), the Virgo observatory near Pisa, Italy, and a similar detector in Japan that could begin making observations next year. They will get an extra boost from space-based interferometers, and from terrestrial ones that are still on the drawing board — as well as from other methods that could soon start producing their own first detections of gravitational waves (see 'The gravitational-wave spectrum').

Like many scientists, Schutz hopes that the best discoveries will be ones that no theorist has even dreamed of. "Any time you start observing something so radically new, there's always the possibility of seeing things you didn't expect."

SPINNING CLUES
For a field of research that is not yet three years old, gravitational-wave astronomy has delivered discoveries at a staggering rate, outpacing even the rosiest expectations. In addition to the discovery in August of the neutron-star merger, LIGO has recorded five pairs of black holes coalescing into larger ones since 2015 (see 'Making waves'). The discoveries are the most direct proof yet that black holes truly exist and have the properties predicted by general relativity. They have also revealed, for the first time, pairs of black holes orbiting each other.

Researchers now hope to find out how such pairings came to be. The individual black holes in each pair should form when massive stars run out of fuel in their cores and collapse, unleashing a supernova explosion and leaving behind a black hole with a mass ranging from a few to a few dozen Suns.

There are two leading scenarios for how such black holes could come to circle each other: they might start as massive stars in each other's orbit, and stay together even after each goes supernova. Or, the black holes might form independently, but be driven together later by frequent gravitational interactions with other objects — something that could happen in the centres of dense star clusters.

Either way, the objects' energy gradually disperses in the form of gravitational waves, a process that pulls the pair into an ever tighter and faster spiral, eventually fusing into one more-massive black hole. Ilya Mandel, a LIGO theorist at the University of Birmingham, UK, says that for LIGO and Virgo to see such pairs merge, typical black holes need to have started their mutual orbit separated by a distance of less than one-quarter that between Earth and the Sun. "If you start out with the two black holes any farther apart, it will take longer than the age of the Universe" for them to merge, Mandel says.
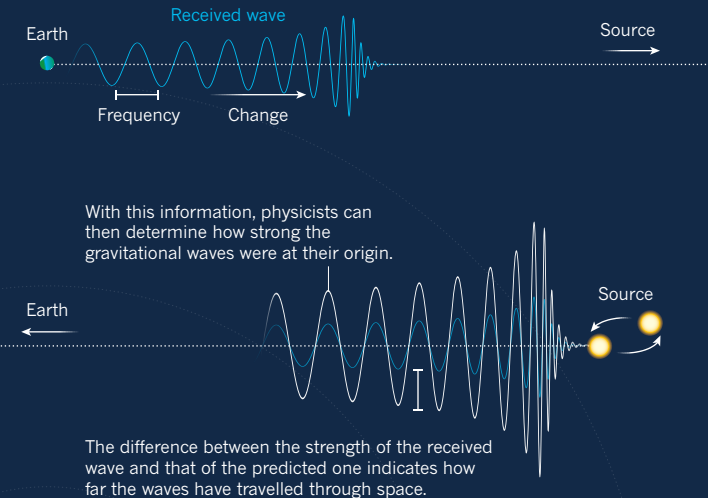
The five black-hole mergers discovered so far are not sufficient to determine which formation scenario dominates. But in an August analysis of the first three detections, a group including Mandel and Will Farr, a theoretical astrophysicist and LIGO member at the University of Birmingham, suggested that just ten more observations could provide substantial evidence in favour of one scenario or the other[1]. This would involve scrutinizing the

MAKING WAVES
When two black holes or neutron stars spiral into each other, they produce distinctive ripples in space-time called gravitational waves. Teams with LIGO's two detectors in the United States and with Virgo, the observatory's counterpart in Italy, have announced the detection of six events so far.
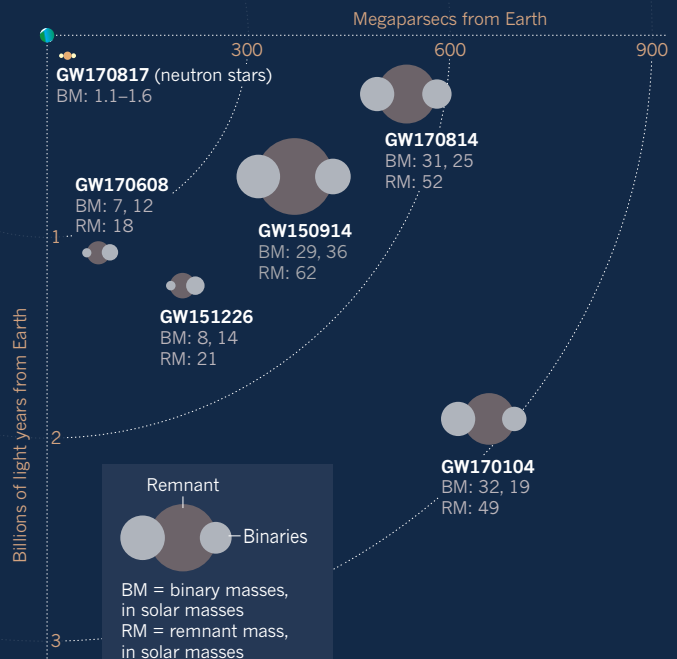
DECIPHERING A WAVE
When a signal is received, the frequency and rate of frequency change provide information about the masses of the objects in the binary source.

Earth — Received wave — Source

Frequency — Change

With this information, physicists can then determine how strong the gravitational waves were at their origin.

Earth — Source

The difference between the strength of the received wave and that of the predicted one indicates how far the waves have travelled through space.

ALREADY DETECTED BY LIGO AND VIRGO
Here are the binary mergers that the observatories have picked up so far. Each discovery was named with the date it was detected.

Megaparsecs from Earth

300   600   900

Billions of light years from Earth

GW170817 (neutron stars)
BM: 1.1–1.6

GW170814
BM: 31, 25
RM: 52

GW170608
BM: 7, 12
RM: 18

GW150914
BM: 29, 36
RM: 62

GW151226
BM: 8, 14
RM: 21

GW170104
BM: 32, 19
RM: 49

Remnant
Binaries

BM = binary masses, in solar masses
RM = remnant mass, in solar masses

gravitational waves for clues about how black holes rotate: those that pair up after forming independently should have randomly oriented spins, whereas those with a common origin should have spin axes that are parallel to each other and roughly perpendicular to the plane in which they orbit.

Further observations could also provide insight into some of the fundamental questions about black-hole formation and stellar

evolution. Collecting many measurements of masses should reveal gaps — ranges in which few or no black holes exist, says Vicky Kalogera, a LIGO astrophysicist at Northwestern University in Evanston, Illinois. In particular, "there should be a paucity of black holes at the low-mass end", she says, because relatively small supernovae tend to leave behind neutron stars, not black holes, as remnants. And at the high end — around 50 times the mass of the Sun — researchers expect to see another cut-off. In very large stars, pressures at the core are thought eventually to produce antimatter, causing an explosion so violent that the star simply disintegrates without leaving any remnants at all. These events, called pair-instability supernovae, have been theorized, but so far there has been scant observational evidence to back them up.

Eventually, the black-hole detections will delineate a map of the Universe in the way galaxy surveys currently do, says Rainer Weiss, a physicist at the Massachusetts Institute of Technology in Cambridge who was the principal designer of LIGO. Once the numbers pile up, "we can actually begin to see the whole Universe in black holes", he says. "Every piece of astrophysics will get something out of that."

To ramp up these observations, LIGO and Virgo have plans to improve their sensitivity, which will reveal not only more events, but also more details about each merger. Among other things,

"WE CAN ACTUALLY BEGIN TO SEE THE WHOLE UNIVERSE IN BLACK HOLES"

physicists are eager to see the detailed 'ringdown' waves that a post-merger black hole emanates as it settles into a spherical shape — an observation that could potentially reveal cracks in the general theory of relativity.

Having more observatories spread around the globe will also be crucial. KAGRA, a detector under construction deep underground in Japan, might start gathering data by late 2019. Its location — and, in particular, its orientation with respect to incoming waves — will complement LIGO's and Virgo's, and enable researchers to nail down the polarization of the gravitational waves, which encodes information about the orientation of the orbital plane and the spin of the spiralling objects. And India is planning to build another observatory in the next decade, made in part with spare components from LIGO.

An even bigger trove of discoveries could come from observing neutron-star mergers. So far, researchers have announced only one such detection, called GW170817. That signal, seen last August, was almost certainly the most intensely studied event in astronomy's history. And it solved a number of long-standing mysteries in one stroke, including the origin of gold and other heavy elements in the Universe[2], as well as the cause of some γ-ray bursts[3].

Further observations could allow scientists to explore the interiors of these objects. Neutron stars are thought to be as dense as matter can possibly be without collapsing into a black hole, but exactly how dense is anybody's guess. No laboratory experiment can study those conditions, and there are dozens of proposals for what happens there. Some theories predict that quarks — the subatomic components that make up protons and neutrons — should break free from each other and roam about, perhaps in superconducting, superfluid states. Others posit that heavier, 'strange' quarks form and become part of exotic cousins of the neutron.

Pinning down the radii of neutron stars might allow physicists to evaluate the theories, because they predict different 'equations

of state' — formulae that link pressure, temperature and density of matter. Such equations determine to what extent matter can be compressed, and so how wide or narrow a neutron star will be for a given mass, and how massive such stars can get.

The 100-second-long signal in August eventually became too high in pitch for LIGO and Virgo to detect, which prevented the observatories from seeing the two neutron stars' final moments, when they should have deformed each other in ways that would have revealed their size and hardness, or resistance to compression. Still, says B. S. Sathyaprakash, a LIGO theoretical physicist at the Pennsylvania State University in University Park, from that one event, "we can rule out equations of state that allow neutron-star sizes larger than 15 kilometres in radius" — a figure that is consistent with other measurements and favours 'softer' matter.

Future detections — and detectors — will give much more detail. Sathyaprakash says that the Einstein Telescope, a possible next-generation observatory dreamed up by a team in Europe, could take physicists far beyond an upper limit. "We want to be able to pin down the radius to the level of 100 metres," he says — a precision that would be astounding, given that these objects are millions of light years away.

SIREN CALLS

Signals similar to GW170817, which was observed through both gravitational waves and light, could have dramatic implications for cosmology. Schutz calculated in 1985 that the frequency, or pitch, of waves from spiralling objects, together with the rate at which that pitch increases, reveals information about the objects' collective mass[4]. That determines how strong their waves should be at the source. By measuring the strength of the waves that reach Earth — the amplitude of the signal actually picked up by interferometers — one can then estimate the distance that the waves have travelled from the source. All other things being equal, a source that is twice as far, for example, will produce a signal half as strong. This type of signal has been dubbed a standard siren, in a nod to a common method of gauging distances in cosmology: stars called standard candles have a well-known brightness, which allows researchers to work out their distance from Earth.

By coupling the distance measurement of GW170817 with an estimate of how fast the galaxies in that region are receding from Earth, Schutz and his collaborators made a new and completely independent estimate of the Hubble constant — the Universe's current rate of expansion (see 'Cosmic signposts'). The result[5], part of a crop of papers released by LIGO, Virgo and some 70 other astronomy teams on 16 October (see go.nature.com/2gbsgnq), "ushers in a new era for both cosmology and astrophysics", says Wendy Freedman, an astronomer at the University of Chicago in Illinois who has made highly precise measurements of the Hubble constant, using time-honoured, but less-direct, techniques.
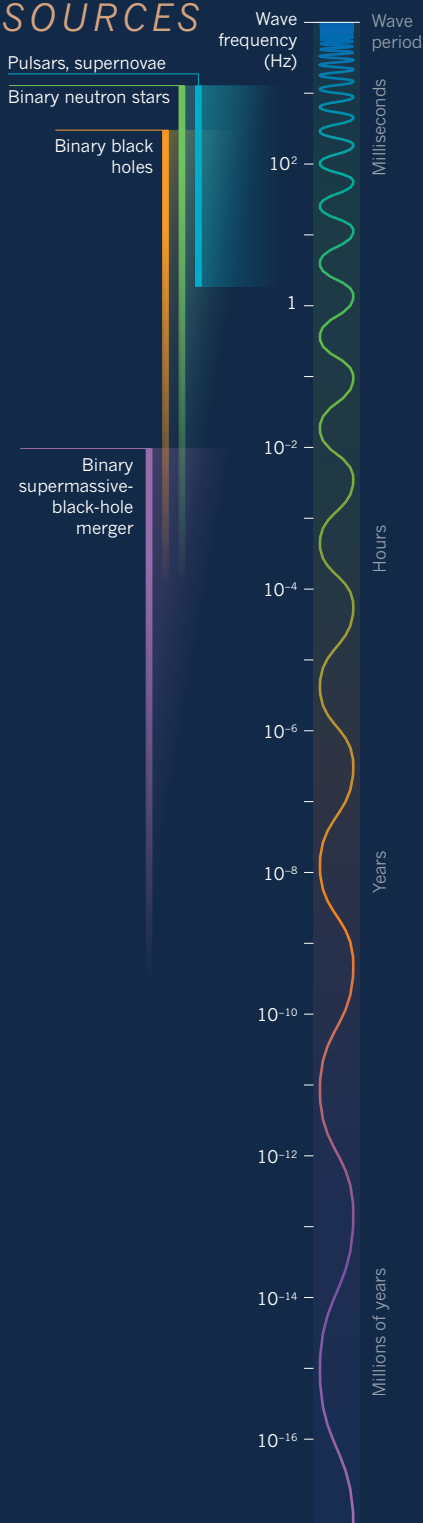
As a direct and independent measure of this constant, standard sirens could help to resolve a disagreement among cosmologists. State-of-the-art techniques, refined over nearly a century of work that started with Edwin Hubble himself, now give estimates that differ by a few per cent. This first standard-siren measurement does not resolve the tension: the expansion rate it predicts falls somewhere in the middle of the range and, because it is based on just one merger event, has a large error bar. But in the future, researchers expect standard sirens to nail down the Hubble constant with an error of less than 1%. So far, standard candles have done it with precisions of 2–3%.

Standard sirens could become even more powerful tools with space-based interferometers such as the Laser Interferometer Space Antenna (LISA), a trio of probes that the European Space Agency, which is leading the mission, plans to launch in the 2030s. LISA is designed to be sensitive to low-frequency waves that ground-based observatories cannot detect. This would give it access to more-massive systems, which radiate stronger

THE GRAVITATIONAL-WAVE SPECTRUM

Much like electromagnetic waves, gravitational waves are emitted by many different objects over a wide range of frequencies. Terrestrial interferometers such as the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo are sensitive to only a subset of those frequencies, which limits their ability to 'see' certain cosmic phenomena. They won't detect collisions of supermassive black holes found in the hearts of galaxies, for example. But space-based interferometers and other approaches for picking up gravitational waves could extend physicists' reach.

SOURCES

Wave frequency (Hz)

Wave period

Pulsars, supernovae
Binary neutron stars
Binary black holes

10^2

1

Milliseconds

Binary supermassive-black-hole merger

10^{-2}

10^{-4}

Hours

10^{-6}

10^{-8}

Years

10^{-10}

10^{-12}

10^{-14}

Millions of years

10^{-16}

DETECTORS

400 Hz – 30 Hz
GROUND-BASED INTERFEROMETER

Current observatories such as LIGO can detect waves that are longer than the detectors' lengths (3–4 kilometres), corresponding to periods of a few hundredths to a few thousandths of a second.

Mirror
Input mirror
Laser
Beam splitter
Light detector
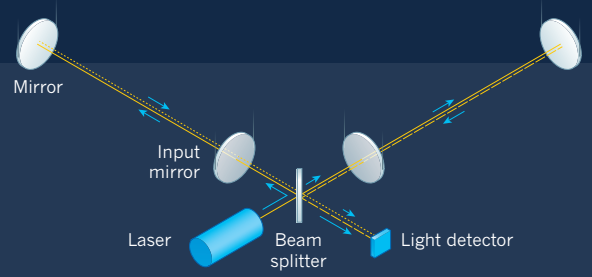
100 milliHz – 0.1 milliHz
SPACE-BASED INTERFEROMETER

LISA, the trio of probes slated to fly in the 2030s, will have virtual arms millions of kilometres long, which will make it sensitive to waves with periods of tens of seconds to a few hours.

Earth
Three spacecraft
Sun
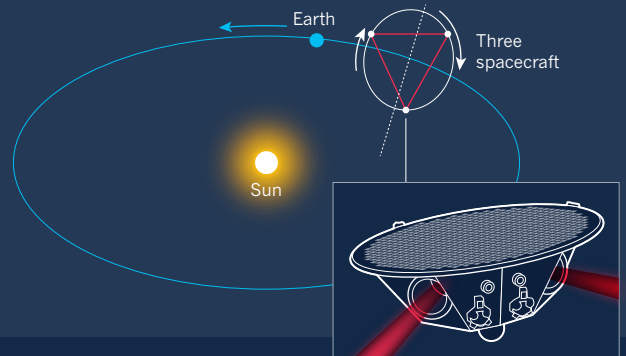
320 nanoHz – 1 nanoHz
PULSAR TIMING

Gravitational waves from distant galaxies perturb the distance between Earth and stars in the Milky Way. Researchers hope to detect waves of periods lasting years, by examining delays in the radio signals from spinning neutron stars known as pulsars.

Pulsar
Radio beam
Earth
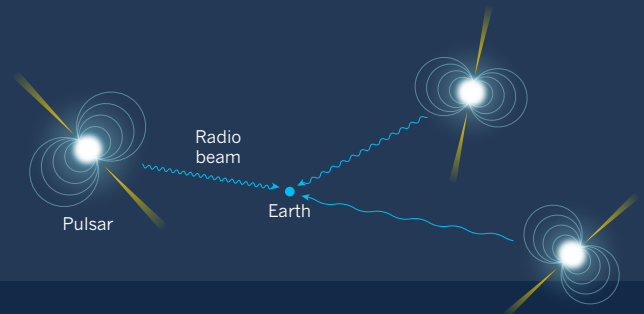
~10^{-13} – 10^{-16} Hz
CMB MEASUREMENT

The Universe's oldest measurable radiation (the cosmic microwave background, or CMB) could carry evidence of gravitational waves from the Big Bang. Those waves would not be detectable more directly; by now, they would stretch across a significant fraction of the observable Universe.

Microwave background
Electromagnetic radiation
Big Bang
Earth
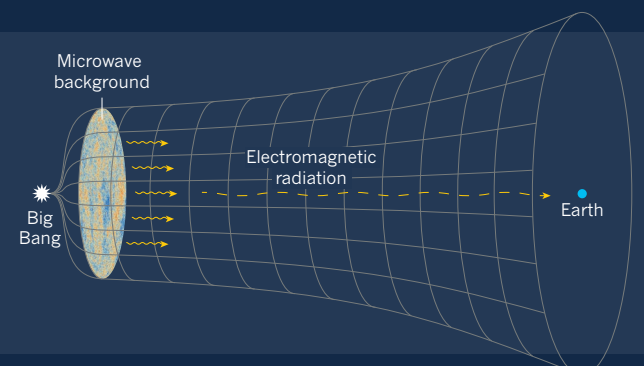
COSMIC SIGNPOSTS

Neutron-star mergers are new tools for measuring the Hubble constant — the current expansion rate of the Universe.

Earth — Gravitational-wave signal — Neutron-star pair — Galaxy

The gravitational-wave signal can be used to gauge the distance from Earth to the former neutron-star pair.

Because the merger event also releases light, conventional telescopes can be used to help pinpoint where it happened.
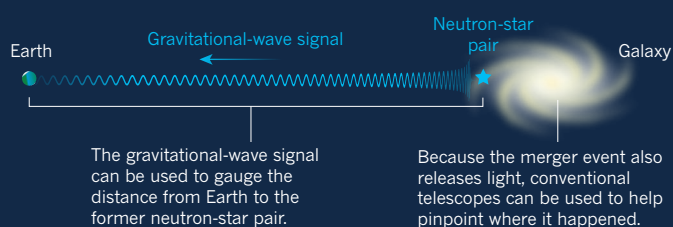
Then, standard astronomical techniques can be used to measure how fast the galaxy and those around it are speeding away from Earth.

Redshifted light

Expanding Universe
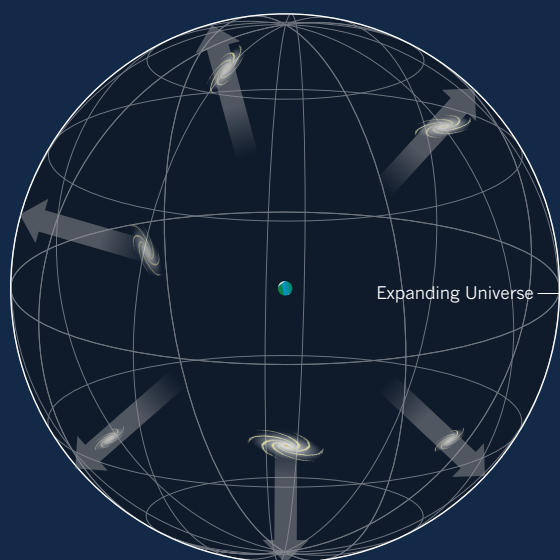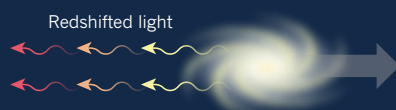
The velocity and distance data — ideally from many such mergers — can be combined to calculate the Hubble constant, which relates distance and speed (galaxies twice as distant recede twice as fast).

NIK SPENCER/NATURE

gravitational waves. In principle, LISA could pick up sirens from across the Universe and, with the help of conventional telescopes, measure not just the current rate of cosmic expansion, but also how that rate has evolved through the aeons. Thus, LISA could help to address cosmology's biggest puzzle: the nature of dark energy, the as-yet-unidentified cosmic component that is driving the Universe's expansion to accelerate.

Whereas ground-based interferometers detect events that are brief and far between, LISA is expected to hear a cacophony of signals as soon as it turns on, including a constant chorus of tight binary white dwarfs — the ubiquitous remnants of Sun-sized stars — in our own galaxy. "It's as if we lived in a noisy forest, and we had to single out the sounds of individual birds," says astrophysicist Monica Colpi of the University of Milan–Bicocca in Italy, who is part of a committee setting the mission's science goals.

Occasionally, LISA should see black-hole mergers such as the

ones LIGO does, but on a much grander scale. Most galaxies are thought to harbour a central supermassive black hole that weighs millions, or even billions, of solar masses. Over a scale of billions of years, galaxies might merge several times; eventually, their central black holes might merge, too. These events are not frequent for individual galaxies, but because there are trillions of galaxies in the observable Universe, a detectable merger should occur somewhere at least a few times per year. Scientists are also pursuing a separate way of detecting gravitational waves from pairs of these behemoths at earlier stages of their orbits. Using radio telescopes, they monitor pulsars inside the Milky Way and look for small variations in their signals, caused by the passage of gravitational waves through the galaxy. Today, there are three 'pulsar-timing arrays', in Australia, Europe and North America, and a fourth forming in China.

Thanks to LISA's planned sensitivity, and the strong signals produced by spiralling supermassive black holes, the observatory should be able to pick up gravitational waves from pairs of supermassive black holes months before they merge, and see the merger in enough detail to test general relativity with high precision. After years of operation, LISA could accumulate enough distant events for researchers to reconstruct the hierarchical formation of galaxies — how small ones combined to form larger and larger ones — in the Universe's history.

On the ground, too, physicists are beginning some "grand new ventures", Weiss says. A US team envisions a Cosmic Explorer with 40-kilometre detecting arms — 10 times as long as LIGO's — that would be sensitive to signals from events much farther away, perhaps across the entire observable Universe.

The concept for the Einstein Telescope in Europe calls for a detector with 10-kilometre arms arranged in an equilateral triangle and placed in tunnels 100 metres or so underground. The quiet conditions there could help to broaden the observatory's reach, to frequencies one-tenth those detectable by current machines. That might allow scientists to find black holes beyond the range thought to be prohibited by pair-instability supernova; at high enough masses, stars should have a different collapse mechanism and be able to form black holes of 100 solar masses or more.

If scientists are lucky, gravitational waves might even let them access the physics of the Big Bang itself, at epochs that are not observable by any other means. In the first instants of the Universe, two fundamental forces — the electromagnetic force and the weak nuclear force — were indistinguishable. When these forces separated, they might have produced gravitational waves that, today, could show up as a "random hiss" detectable by LISA, Schutz says. This hypothetical signal is distinct from a much longer-wavelength one from even earlier on, which might appear in the Universe's oldest visible radiation: the cosmic microwave background. In 2014, a team reported[6] that it had observed this effect with the BICEP2 telescope at the South Pole, but the researchers later acknowledged problems with that interpretation[7].

With the reopening of both LIGO and Virgo late this year, the next big discovery on Weiss's wish list is the signal from a collapsing star — something that astronomers might also observe as a type of supernova. But he has high hopes for what else might be on the horizon. "If we don't see something that we hadn't thought of," Weiss says. "I'd be disappointed." ∎

Davide Castelvecchi *is a senior reporter for* Nature *in London.*

1. Farr, W. M. *et al. Nature* **548**, 426–429 (2017).
2. Smartt, S. J. *et al. Nature* **551**, 75–79 (2017).
3. Goldstein, A. *et al. Astrophys. J. Lett.* **848**, L14 (2017).
4. Schutz, B. F. *Nature* **323**, 310–311 (1986).
5. The LIGO Scientific Collaboration and The Virgo Collaboration *et al. Nature* **551**, 85–88 (2017).
6. Ade, P. A. R. *et al. Phys. Rev. Lett.* **112**, 241101 (2014).
7. Ade, P. A. R. *et al. Phys. Rev. Lett.* **114**, 101301 (2015).

COMMENT

A safety driver sits behind the wheel during a test of a self-driving taxi in Yokohama, Japan.

People must retain control of autonomous vehicles

Legislation on the testing of self-driving cars does not address liability and safety concerns, warn **Ashley Nunes**, **Bryan Reimer** and **Joseph F. Coughlin**.

Last month, for the first time, a pedestrian was killed in an accident involving a self-driving car. A sports-utility vehicle controlled by an autonomous algorithm hit a woman who was crossing the road in Tempe, Arizona. The safety driver inside the vehicle was unable to prevent the crash.

Although such accidents are rare, their incidence could rise as more vehicles that are capable of driving without human intervention are tested on public roads. In the past year, several countries have passed laws to pave the way for such trials. For example, Singapore modified its Road Traffic Act to permit autonomous cars to drive in designated areas. The Swedish Transport Agency allowed driverless buses to run in northern Stockholm. In the United States, the House of Representatives passed the SELF DRIVE Act to harmonize laws across various states. Similar action is pending in the US Senate, where a vote to support the AV START Act would further liberalize trials of driverless vehicles.

Policymakers are enthusiastic about the potential of autonomous vehicles to reduce road congestion, air pollution and road-traffic accidents[1,2]. Cheap ride-hailing services could reduce the number of privately owned cars. Machine intelligence can make driving more fuel-efficient, cutting emissions. Autonomous cars could help to save the 1.25 million lives worldwide that are lost each year through crashes[3], many of which are caused by human error.

Governments want to pass laws to make this happen (see 'Road to autonomy'). But they are doing so by temporarily freeing developers of self-driving cars from ▶

meeting certain transport safety rules. These rules include the requirement that a human operator be inside the vehicle, that vehicles have safety features such as a steering wheel, brakes and a mirror, and that the features are functional at all times. Some developers are maintaining these aspects, but they are not obliged to do so. There is no guarantee that autonomous vehicles will match the safety standards of current cars.

Meanwhile, the wider policy implications are not being addressed[1,2]. Governments stand to lose billions of dollars in tax revenue as rates of car ownership drop among individuals. Millions of taxi, lorry and bus drivers will lose their jobs[2]. The machine-learning algorithms on which autonomous vehicles rely are far from developed enough to make choices that could mean life or death for pedestrians or drivers.

Policymakers need to work more closely with academics and manufacturers to design appropriate regulations. This is extremely challenging because the research cuts across many disciplines.

Here, we highlight two areas — liability and safety — that require urgent attention.

LIABILITY

Like other producers, developers of autonomous vehicles are legally liable for damages that stem from the defective design, manufacture and marketing of their products. The potential liability risk is great for driverless cars because complex systems interact in ways that are unexpected.

Manufacturers want to minimize the number of liability claims made against them[4]. One way is to reduce the chance of their product being misused by educating consumers about how it works and alerting them to safety concerns. For example, drug developers provide information on dosages and side effects; electronics manufacturers

issue instructions and warnings. Such guidance shapes the expectations of consumers and fosters satisfaction. Yet, much like smartphones, self-driving cars are underpinned by sophisticated technologies that are hard to explain or understand.

Instead, developers are designing such products to be easy to use[5]. People are more likely to buy a product that seems straightforward and with which they can soon do complicated things, increasing its utility. However, users are then less able to anticipate how the underlying systems work, or to recognize problems and fix them. For example, few drivers of computerized cars know how the engine is calibrated[5]. Similarly, a passenger in an autonomous vehicle will not know why it chooses to make a sharp turn into oncoming traffic or why it does not overtake a slow-moving vehicle.

Worse, deep-learning algorithms are inherently unpredictable. They are built on an opaque decision-making process that is shaped by previous experiences. Each car will be trained differently. No one — not even an algorithm's designer — can know precisely how an autonomous car will behave under every circumstance.

No law specifies how much training is needed before a deep-learning car can be deemed safe, nor what that training should be. Cars from different manufacturers could react in contrasting ways in an emergency. One might swerve around an obstacle; another might slam on the brakes. Rare traffic events, such as a truck tipping over in the wind, are of particular concern and, at best, make it difficult to train driverless cars.

Advanced interfaces are needed that inform users why an autonomous vehicle is behaving as it does. Today's dashboards convey information about a car's speed and the amount of fuel that remains. Tomorrow's displays must show the vehicle's 'intentions'

and the logic that governs them; for example, they might tell passengers that the car will not overtake the vehicle ahead because there is only a 10% likelihood of success. Little is known about the types of data that should be imparted and how people will interpret them.

Users often ignore information, even if it is presented clearly and the consequences could be a matter of life or death. For instance, almost 70% of airline passengers do not review safety cards before a flight[6], despite being asked. Yet these cards convey important information, including how to put on an oxygen mask and open an emergency exit, in simple terms and on a single page.

Autonomous vehicles will need to communicate much more complicated information. Their sensors and algorithms must understand the behaviours of pedestrians, discriminate between styles of driving and adjust to changes in lighting. When they cannot, users must know how to respond.

Researching ways to present this information effectively is paramount, as are legislative efforts to ensure that users of autonomous vehicles are proficient in using the technology.

SAFETY

The safety and efficiency benefits of autonomous cars rely on computers making better, quicker decisions than people. Users input their desired destination and thereafter cede control to the computer. Full autonomy has — deliberately — not yet been adopted in transportation. People are still perceived as being more flexible, adaptable and creative than machines, and better able to respond to changing or unforeseen conditions[7]. Pilots are able, therefore, to wrest control from fly-by-wire technology when key computers fail.

The public is right to remain cautious about full automation. Manufacturers need to explain how a car would protect passengers should crucial systems fail. A driverless car must be able to stop safely if its hazard-avoidance algorithms malfunction, its cameras break or its internal maps die. But this is hard to engineer: for example, without cameras, such a car cannot see where it is going.

In our view, some form of human intervention will always be required. Driverless cars should be treated much like aircraft, in which the involvement of people is required despite such systems being highly automated. Current testing of autonomous vehicles abides by this principle. Safety drivers are present, even though developers and regulators talk of full automation.

Nonetheless, having people involved poses safety problems. Autonomous cars will always require users to have a minimum level of skill and will never be easy for some members of the public to operate. People with cognitive impairments, say, might find it difficult to operate these technologies and to override controls. Yet this group includes those who would benefit greatly
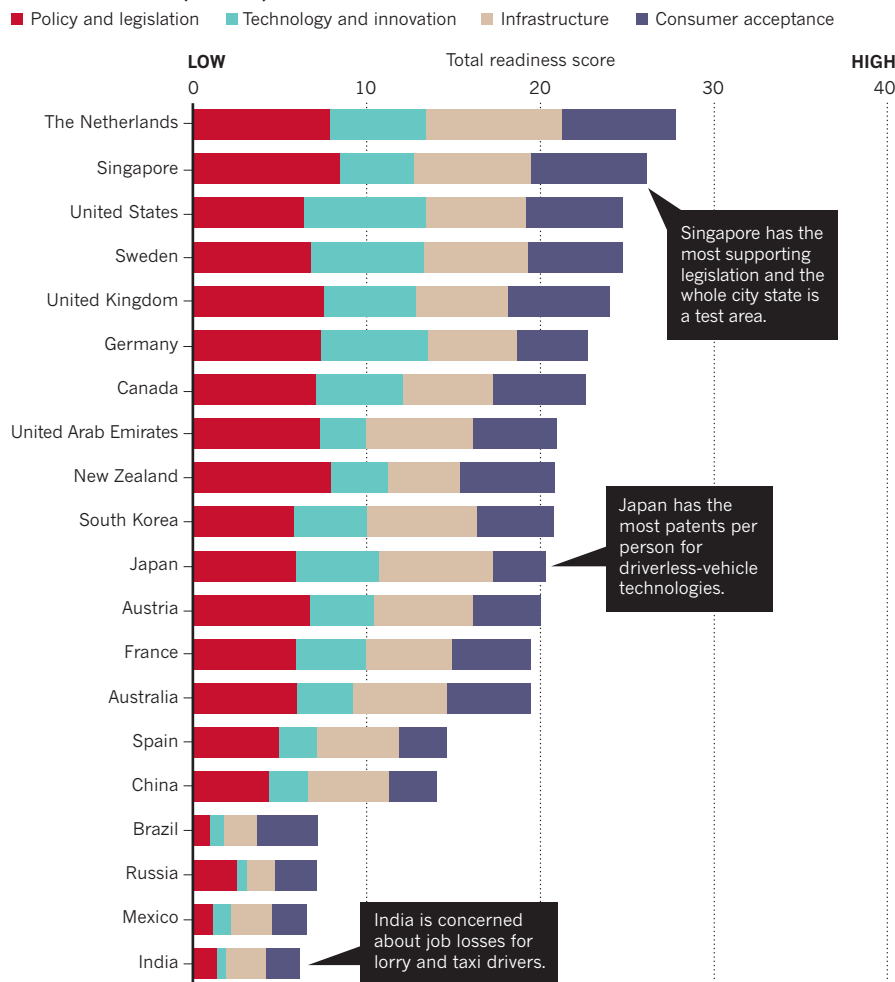

A driverless bus shuttles passengers across Southeast University's Jiulonghu campus in Nanjing, China.

VCG/GETTY

ROAD TO AUTONOMY
The Netherlands heads the list of countries that are most prepared for autonomous vehicles. Twenty nations were assessed according to four key areas of preparedness.

READINESS SCORE (out of 10)
- Policy and legislation
- Technology and innovation
- Infrastructure
- Consumer acceptance

SOURCE: AUTONOMOUS VEHICLES READINESS INDEX (KPMG INTERNATIONAL, 2018)

Total readiness score — LOW 0 — 10 — 20 — 30 — HIGH 40

The Netherlands, Singapore, United States, Sweden, United Kingdom, Germany, Canada, United Arab Emirates, New Zealand, South Korea, Japan, Austria, France, Australia, Spain, China, Brazil, Russia, Mexico, India

Singapore has the most supporting legislation and the whole city state is a test area.

Japan has the most patents per person for driverless-vehicle technologies.

India is concerned about job losses for lorry and taxi drivers.

Users need information on how autonomous systems are working. Manufacturers must research the limits and reliability of devices that are crucial for safety, including cameras, lasers and radars. When possible, they should make the data from these devices available to vehicle operators in an understandable form.

Operators must demonstrate competence. Developers, researchers and regulators need to agree proficiency standards for users of autonomous vehicles. Competency should be tested by licensing authorities and should supplement existing driving permits. Users who fall short should have their access to such vehicles limited, just as colour-blind pilots are banned from flying at night.

Regular checks on user competency should be mandatory. Regulators, manufacturers and researchers must determine a suitable time interval between tests, so that proficiency is kept up as cognitive abilities change and technology evolves.

Remote monitoring networks should be established. Manufacturers, researchers and legislators need to build supervisory systems for autonomous vehicles. Researchers should supply guidance on the number of vehicles that one supervisor can monitor safely, and on the conditions under which such monitoring is permissible. For example, more supervisors would be needed in poor weather conditions.

Work limits for remote supervisors should be defined. Experts must clarify whether supervisors should be subject to existing working-time regulations. For example, air-traffic controllers are limited in how long they can work.

The path towards autonomy is far from preordained. Considerable challenges remain to be addressed. ∎

Ashley Nunes, Bryan Reimer and Joseph F. Coughlin are in the AgeLab, Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.
e-mail: anunes@mit.edu

1. Fagnant, D. J. & Kockelman, K. Transport. Res. Part A Policy Pract. 77, 167–181 (2015).
2. Anderson, J. M. et al. Autonomous Vehicle Technology: A Guide for Policymakers (Rand Corporation, 2016).
3. World Health Organization. Global Status Report on Road Safety: Time for Action (WHO, 2015).
4. Henderson, J. A. Jr. Ariz. State Law J. 47, 1145–1180 (2015).
5. Frazzini, R. IEEE Control Systems 21, 8–12 (2001).
6. National Transportation Safety Board. Emergency Evacuation of Commercial Airplanes. Safety Study NTSB/SS-00/01 (NTSB, 2000).
7. Parasuraman, R. & Riley, V. Hum. Factors 39, 230–253 (1997).
8. Christensen, K., Doblhammer, G., Rau, R. & Vaupel, J. W. Lancet 374, 1196–1208 (2009).
9. National Highway Traffic Safety Administration. Traffic Safety Facts: Older Population. Report No. DOT HS 812 005 (NHTSA, 2012).
10. Owsley, C. et al. JAMA 279, 1083–1038 (1998).
11. Kahneman, D. Attention and Effort Ch. 9, 136–155 (Prentice Hall, 1973).

from self-driving vehicles. For example, older adults[8], a demographic of increasing importance, have an elevated risk of crashes because cognitive abilities decline with age[9,10]. Providing mobility for large numbers of elderly people is an impetus for investment in this technology in Japan, for instance.

A remote supervisor could oversee driverless cars as air-traffic controllers do for aircraft. But how many supervisors would be needed to keep networks of such vehicles safe? Stretching human capacity too far can create accidents[11]. For example, in 1991, an overwhelmed air-traffic controller in Los Angeles, California, mistakenly cleared an aeroplane to land on another. Last year, an overload of patients was blamed for a string of medical errors by doctors in Hong Kong.

POLICY GAPS
Current and planned legislation fails to address these issues. Exempting developers from safety rules poses risks. And developers are not always required to report system failures or to establish competency standards for vehicle operators. Such exemptions also presume, wrongly, that human involvement will ultimately be unnecessary. Favouring industry over users will erode support for the technology from an already sceptical public.

Present legislation sidesteps the education of consumers. The US acts merely require that users are "informed" about the technology before its use. Standards of competency and regular proficiency testing for users are not mentioned. Without standards, it is hard to tell whether consumer education programmes are adequate. And without testing, the risk of incidents might increase.

MOVING FORWARD
We call on policymakers to rethink their approach to regulating autonomous vehicles and to consider the following six points when drafting legislation.

Driverless does not, and should not, mean without a human operator. Regulators and manufacturers must acknowledge, rather, that automation changes the nature of the work that people perform[7].

Contemplating the night sky in Maine became a metaphysical experience for physicist Alan Lightman.

A physicist faces the sublime

Anil Ananthaswamy on Alan Lightman's journeys in empiricism and experience.

With his debut novel, *Einstein's Dreams* (1992) — the poetic musings of a Swiss patent clerk on the nature of time — theoretical physicist Alan Lightman revealed an enthusiasm for entering the human psyche. His latest book, the collection of essays *Searching for Stars on an Island in Maine*, goes further. Here, Lightman confronts the contradictions that arise from having a rigorous scientific world view, alongside his own mortal desires and fears.

Lightman begins *Searching for Stars* with an account of a mystical experience. He's motoring through the coastal waters off mainland Maine, towards Pole Island, where he has a summer home. It's a moonless night. Before he docks, Lightman turns off the boat's motor and running lights, and lies down in the silence and darkness. "After a few minutes, my world had dissolved into that star-littered sky. The boat disappeared. My body disappeared ... I felt connected not only to the stars but to all of nature, and to the entire cosmos." With that, Lightman begins an exploration of the tensions, both

within himself and without, between the materialist reductionism of science, especially physics, and the absolutes of spiritual belief.

As a physicist, he knows there are no absolutes. The idea of a fixed and motionless Earth was disproved in 1851 by the "slow rotation of the plane of a swinging pendulum" — physicist Léon Foucault's experiment — which could be explained only if the planet, not the pendulum, was rotating. Discoveries of the electron and radioactivity showed us that even atoms, once thought indestructible, were anything but. Next, Albert Einstein demolished Newtonian notions of absolute space and time. Then came quantum mechanics, with its claims of uncertainty and indeterminism.
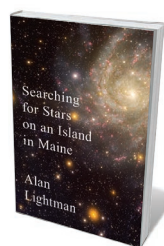
Searching for Stars on an Island in Maine
ALAN LIGHTMAN
Pantheon: 2018

For anyone looking to science for assurance, the bottom falls out. So, Lightman looks past it. "I am a scientist, but I am not a swinging bob on a string," he writes.

He sets the stage for a dialogue, introducing us to the usual suspects in science: Galileo Galilei, J. J. Thomson, Ernest Rutherford and Einstein; and to a handful of spiritual thinkers and religious figures. We meet, for instance, fleetingly, the Indian poet and Nobel laureate in literature Rabindranath Tagore; and, more substantially, Augustine of Hippo, the influential fifth-century Christian theologian.

"Augustine's certainties were absolute," Lightman writes, contrasting these immutable religious ideas — such as the immortal soul — with science's ever-evolving view. Yet, he argues, science, too, longs for an absolute in a final 'theory of everything', and has its article of faith: "that the physical world is a territory of order and logic".

Lightman's scope is sweeping, but he doesn't dig deeply enough. For instance, he expresses disbelief in *bardo* — the Tibetan

Buddhist term signifying the transitory state between death and rebirth. He writes: "I ask for some kind of evidence for all things I believe — even if it is evidence from a personal or transcendent experience. And I insist on evidence for any statements that concern the physical world." Certainly, there is no 'evidence' for *bardo*, independent of the subjective experiences of Tibetan Buddhists. But Lightman stands by his own subjective experience of perceived 'oneness' with something larger than ourselves. A rigorous scientific approach would question the veracity of all subjective experiences, not just those that seem unreasonable at first blush.

> *"As a physicist, Lightman knows there are no absolutes."*

Therein lies the book's Achilles heel: it makes little mention of the research on perception that calls into doubt the 'truth' of subjective experiences, no matter how real or exalted they feel. Modern neuroscience tells us that what we perceive is not a bottom-up reconstruction by the brain of what's out there. Rather, it is the brain's prediction about the probable causes of sensory inputs.

Predictions, and thus perceptions, can be wrong. This is of particular importance when perceptions hint at something spiritual. It's a subject explored, for example, in the 2015 *Kabbalah: A Neurocognitive Approach to Mystical Experiences* by neurologist Shahar Arzy and scholar of Jewish thought Moshe Idel. Mystical experiences are also eerily similar to those reported by people having ecstatic epileptic seizures, including feelings of time dilation and 'oneness', the neural underpinnings of which are under study (M. Gschwind and F. Picard *Front. Behav. Neurosci.* **10,** 21; 2016).

The book's narrative structure — set up as the articulate reveries of a physicist, who is alternately naturalist, stargazer and philosopher, wandering around his island, constantly thinking grand thoughts on mossy slopes — risks becoming self-parody. Lightman saves the day somewhat by acknowledging the indulgence.

However, as a broad take on intellectual thought at the intersection of science and spirituality, *Searching for Stars* is stimulating. Lightman is to be admired for his willingness to take off his scientist's hat and plunge into preoccupations most of his peers would strenuously avoid, some for fear of ridicule. Once again, this deft wordsmith has effortlessly straddled the divide between the hardest of the hard sciences and the nebulous world of existential doubts and longings. ∎

Anil Ananthaswamy *is a journalist and author of* The Man Who Wasn't There, *an exploration of the neuroscience of the sense of self.*
e-mail: anil@nasw.org

Books in brief
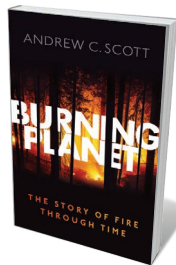
Burning Planet
Andrew C. Scott OXFORD UNIVERSITY PRESS *(2018)*
Megafires regularly crackle through the headlines, yet wildfire management remains largely misguided. Geologist Andrew Scott redresses the balance in this scholarly yet accessible study, drawing on ground and satellite observation as well as his original research into the 400-million-year history of fire on Earth. Through technologies such as scanning electron microscopy, Scott's study of fossil charcoal has unearthed an astounding deep-past record of botanical riches and shifts in climate and oxygen levels. A timely book in an era of heightened fire risk and threats to water supply.
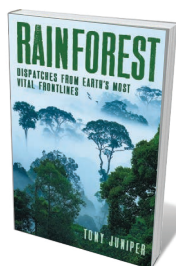
Rainforest: Dispatches from Earth's Most Vital Frontlines
*Tony Juniper* PROFILE *(2018)*
The "green oceans" that are tropical rainforests help to regulate Earth's water, climate and carbon cycles; support 50% of terrestrial flora and fauna; and offer a lifeline to 1.6 billion people. Yet half have been cleared, in large part by consumer-led interests, from cattle ranching to palm-oil production. Environmentalist Tony Juniper surveys the terrain through myriad lenses: the bitter history of exploitation and its impact on indigenous peoples; the stupendous biological riches; and the conservation science and community involvement that, given political and industrial will, could halt the felling.
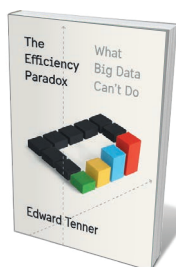
The Efficiency Paradox: What Big Data Can't Do
*Edward Tenner* KNOPF *(2018)*
We pursue efficiency through engulfment in the digital. Yet, argues historian of technology Edward Tenner in this perceptive study, the promise of big data and algorithms for information, education, medicine and beyond is dissipating. The Silicon Valley dream of a frictionless existence is failing because ethical, political and social elements were factored in poorly, spawning issues such as flawed algorithms. Sympathetically critiquing the work of others in this arena, including Nicholas Carr and Cathy O'Neill, Tenner calls for a strategy that blends intuition and experience with high technology.
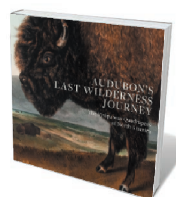
Audubon's Last Wilderness Journey
Marilyn Laufer et al. GILES *(2018)*
Forget birds: otters, bison, armadillos, black bears, elk, beavers and other New World mammals starred in ornithologist John James Audubon's last great work of natural-history illustration. Published in three volumes between 1845 and 1848, and inspired by Audubon's 1843 journey up the Missouri River, the original featured 150 hand-coloured illustrations. Curators at the Jule Collins Smith Museum of Fine Art at Alabama's Auburn University have now made them available to all. Accompanying the striking reproductions are fresh essays on hunting, conservation, wilderness, mammalogy and more.
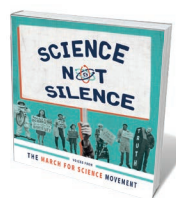
Science Not Silence
Edited by Stephanie Fine Sasse and Lucky Tran MIT PRESS *(2018)*
More than one million researchers, postdocs and science aficionados took to the streets across some 600 cities on 22 April 2017. The March for Science was a riposte to the US administration's ennui around research; it aimed to reify the fundamental, multidimensional importance of science in tackling global challenges and advancing knowledge. This vibrant photo-essay compilation, edited by science communicators Stephanie Fine Sasse and Lucky Tran, pays homage to the international community and its resilience, creativity and ongoing commitment to speaking truth to power. **Barbara Kiser**

Correspondence

Don't dismiss non-English citations

We find it inexcusable for peer reviewers to dismiss citations to scientific papers that are not published in English. Journals written in other languages are a valuable repository for much locally relevant applied science (see, for example, M. Neff *Nature* **554**, 169; 2018). And in most countries today, these works are accessible through free, automated translation services.

We experienced such discrimination after submitting a paper to an English-language journal. It was a bibliometric evaluation of research activities at universities in Belarus and Ukraine, so some citations were inevitably in Russian. One reviewer complained that this "precludes … checking that source to determine if it does actually support the authors' statements". Another demanded more information in the text about the work of an internationally recognized bibliometrician, Irina Marshakova-Shaikevich, "since she writes in Russian".

In our view, substituting non-English citations with anglophone alternatives risks transposing credit for ideas and violates citation standards. Papers should be evaluated on academic criteria, not on superficial grounds of communication.

Vladimir S. Lazarev *Belarusian National Technical University, Minsk, Belarus.*

Serhii A. Nazarovets *Kiev National University of Culture and Arts, Kiev, Ukraine.*
vslazarev@bntu.by

Building rapport for better policymaking

Marie Claire Brisbois and colleagues advise scientists to interact with government policy analysts to improve evidence-based policy (see *Nature* **555**, 165; 2018). In our experience, the interaction between academia and policymakers needs to be a two-way process.

We are members of the Centre for the Evaluation of Complexity across the Nexus, a consortium of academics and practitioners who work with UK government departments and agencies to improve policy evaluation and design across the water–energy–food–environment nexus (www.cecan.ac.uk). We test and promote innovative methods and approaches through co-designed and co-produced case studies that span, for instance, rural policy after Brexit, energy security and food-safety regulation.

Progress in these complex policy areas depends on sharing knowledge and building trust and capacity with civil servants across the political spectrum.

Adam P. Hejnowicz, Sue E. Hartley *University of York, UK.*
Nigel Gilbert *University of Surrey, Guildford, UK.*
adam.hejnowicz@york.ac.uk

World Heritage Site fish faces extinction

The North Sea houting (*Coregonus oxyrinchus*) is a whitefish that is endemic to the Wadden Sea, an area including the North Sea coasts of the Netherlands, Germany and Denmark. A critically small population in Denmark's Vidaa River, estimated at 3,500 adult individuals in 2014, is the last remaining worldwide. We call on the Danish authorities to prevent further decline of this fish through informed conservation planning and management before it is too late.

The Wadden Sea is a World Heritage Site that harbours the world's largest unbroken system of intertidal sand and mud flats. The North Sea houting is protected under the Bern Convention and the EU Habitats Directive. Yet Denmark's conservation efforts since 1992 have been limited to population estimates, insufficient regulation of the predatory great cormorant (*Phalacrocorax carbo*) and unsuccessful habitat-restoration projects.

The habitats needed by this fish for spawning and juvenile development are still unclear, so it is not possible to protect or restore them. This basic knowledge is essential for future restoration projects. We urgently need to understand why the population is still in decline and to put effective conservation measures in place. The North Sea houting must not end up next to the great auk (*Pinguinus impennis*) on museum shelves.

Jon C. Svendsen *Technical University of Denmark, Kongens Lyngby, Denmark.*
Aage K. O. Alstrup *Aarhus University, Aarhus, Denmark.*
Lasse F. Jensen *Aalborg University, Aalborg, Denmark.*
jos@aqua.dtu.dk

Pesticide policies need holistic view

New pesticide policies are needed for more sustainable agricultural production, but their wider implications need to be considered. Efforts to ban ubiquitous pesticides such as glyphosate and neonicotinoids are ongoing (see, for example, *Nature* **555**, 150–151; 2018).

In Switzerland, proposals have been made to suspend subsidies for farms that use pesticides and to ban all synthetic pesticides. In Italy, the municipality Mals has banned pesticide use by farmers. Furthermore, private industries are increasingly restricting pesticides and have introduced labels for glyphosate-free products.

Stricter policies can have unintended effects, however. They may encourage changes in land use and management practices that decrease food production and quality, or increase soil erosion and greenhouse-gas emissions. Banned pesticides might even be substituted with more harmful ones.

Technologies such as sensors, drones and robots could help to monitor and control pesticide application (see A. Walter *et al. Proc. Natl Acad. Sci. USA* **114**, 6148-6150; 2017). Pesticide taxation is another, complementary possibility (R. Finger *et al. Ecol. Econ.* **134**, 263–266; 2017).

To avoid misguided policies, trade-offs between different policy goals need to be quantified for a holistic assessment. For example, modelling approaches could assess the impact of more-stringent pesticide policies on plant protection and land use and quantify the economic consequences (T. Böcker *et al. Ecol. Econ.* **145**, 182–193; 2018).

Robert Finger *ETH Zurich, Switzerland.*
rofinger@ethz.ch

Singapore Index for climate change

The Singapore Index of Cities' Biodiversity was set up ten years ago by the National Parks Board of Singapore and the United Nations Secretariat of the Convention on Biological Diversity as urban development boomed. I suggest that this self-assessment tool could also be applied to safeguard biodiversity against the effects of climate change on cities (see X. Bai *et al. Nature* **555**, 23–25; 2018).

The index consolidates important biodiversity indicators to help cities to evaluate and benchmark their conservation efforts (see go.nature.com/2hammaa). The National Parks Board of Singapore (see www.nparks.gov.sg) received the 2017 UNESCO Sultan Qaboos Prize for Environmental Preservation, and the board's experience could benefit cities across the globe.

Singapore should continue to apply the index to manage biodiversity in the face of climate change, for example in heat islands or in areas prone to flooding.

Sameen Ahmed Khan *Dhofar University, Salalah, Oman.*
rohelakhan@yahoo.com

NEWS & VIEWS

The certainty of randomness

Communication systems rely on random-number generators for the encryption of information. A method for producing truly random numbers even from untrustworthy devices could lead to improvements in security. SEE LETTER P.223

STEFANO PIRONIO

Encryption schemes used in modern cryptography make extensive use of random, unpredictable numbers to ensure that an adversary cannot decipher encrypted data or messages. Reliable random-number generators are therefore crucial. For instance, an Internet-wide analysis identified tens of thousands of servers that are vulnerable to basic attacks because of the use of poor-quality random-number generators[1]. On page 223, Bierhorst et al.[2] exploit effects at the crossroads of quantum physics and special relativity to demonstrate the ultimate random-number generator, achieving unprecedented security.

Although schemes to generate random-looking numbers are easy to come up with, assessing their security — the extent to which they are truly unpredictable to a potential adversary — is notoriously difficult. Much of the trouble stems from the fact that such schemes cannot be tested by merely looking at their output from a black-box perspective: that is, a perspective from which the internal workings are unknown. For instance, certain arithmetic operations known as pseudorandom number generators produce sequences of numbers that are completely predictable. However, these sequences do not have any recognizable patterns and thus, from the perspective of someone who does not know how the numbers have been generated, they cannot easily be distinguished from sequences obtained by truly random methods.

It would therefore seem that security can be established only if the random-number generator satisfies two conditions. First, the user must know how the numbers have been generated to verify that a valid procedure is being implemented. And second, the system must be a black box from the adversary's perspective to prevent them from exploiting knowledge about its internal mechanism.

However, the first condition is unrealistic. A random-number generator can deviate from its intended design because of imperfections, component ageing, accidental failures or explicit tampering by an adversary, leading to undetected biases. And monitoring the internal mechanism of a random-number generator in real time is both impractical and
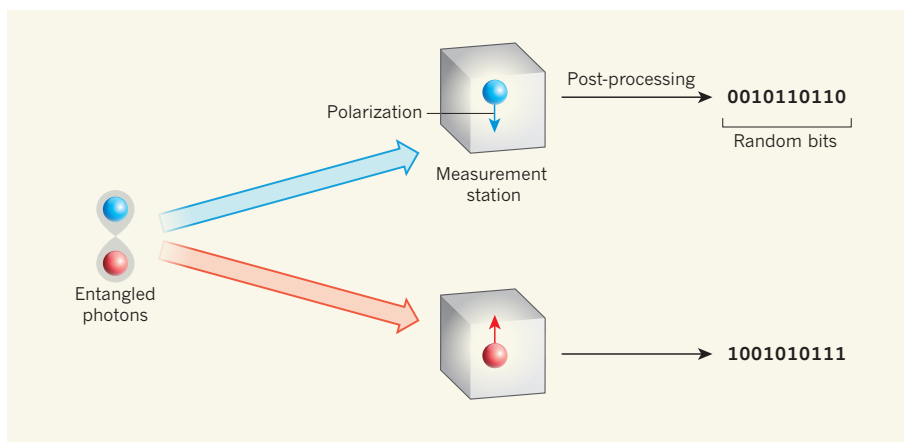
Figure 1 | A quantum random-number generator. Bierhorst et al.[2] report an experiment that produces strings of truly random bits (0s and 1s), which are desirable for improving the security of a wide range of communication systems. The authors prepared a pair of photons (blue and red) that were entangled, meaning that their properties were strongly correlated. They then sent each photon to a different remote measurement station, where the photons' polarizations were recorded. The measurement outcomes from the two stations were unpredictable, thanks to the strong correlated behaviour and large separation of the photons. However, the randomness was small, even after millions of runs. The authors used a powerful post-processing technique to generate truly random bits from these measurements, with minimal physical assumptions about the photons' behaviour.

difficult[3]. Moreover, the second condition violates Kerckhoffs's principle — a central tenet of modern cryptography that was reformulated by the father of information theory, Claude Shannon[4], as "the enemy knows the system being used". In other words, cryptographic systems should be designed under the assumption that an adversary will quickly gain familiarity with them.

Remarkably, thanks to the unusual laws of quantum physics, it is possible to create a provably secure random-number generator for which the user has no knowledge about the internal generation mechanism, whereas the adversary has a fully detailed description of it. To understand how this works, consider the experiment carried out by Bierhorst and colleagues (Fig. 1). The authors prepared two photons in a peculiar quantum condition known as an entangled state. They then sent each photon to a different remote measurement station, where the photons' polarizations were recorded. During measurement, the photons were unable to interact with each other — the stations were so distant that this would require signals travelling faster than the speed of light. Nevertheless, the measurement

outcomes were strongly correlated because of the photons' entangled nature. Such correlations can be detected experimentally through statistical criteria known as violations of Bell inequalities[5].

The strong correlated behaviour of the two remote photons suggests that they could be used to devise a faster-than-light communication device. This would indeed be possible unless the photons' measurement outcomes were unpredictable, in which case any attempt to use such photons in a communication device would fail, because it would result in scrambled, indecipherable messages. Because faster-than-light communication is impossible, it follows that violations of Bell inequalities imply random measurement outcomes. That is, the violations provide an experimental signature of randomness.

This conclusion depends only on the impossibility of faster-than-light signalling and not on any detailed description of the associated quantum systems. It must therefore be true from an adversary's perspective, regardless of their particular knowledge of the quantum processes being carried out. And because violations of Bell inequalities can be verified by

a user only from the statistics of the observed outputs of such processes, the verification procedure represents a black-box test of randomness.

Violations of Bell inequalities have been observed in numerous experiments over the past three decades[5], and their qualitative connection to randomness has been known for many years. However, quantum-information researchers have started to develop the tools to exploit this connection only in the past few years[6].

A key difficulty has been that most experiments that violate Bell inequalities are affected by loopholes, meaning that they cannot be considered as black-box demonstrations. For instance, the constraint that the two photons cannot exchange signals at subluminal speeds was not strictly enforced in the two previous demonstrations of randomness generation based on Bell inequalities[7,8]. In the past few years, loophole-free experiments have been carried out[9–11], but they remain a technological challenge. In particular, the magnitude of the Bell-inequality violations observed in these experiments, although sufficient to confirm the correlated behaviour of the photons, was too low to verify the presence of randomness of sufficient quality for cryptographic purposes.

Bierhorst and co-workers have improved existing loophole-free experimental set-ups to the point at which the realization of such randomness becomes possible. However, this threshold is barely reached. Every time a photon is measured in the authors' experiment, the randomness that is generated (expressed as bits; 0s and 1s) is equivalent to tossing a coin that has 99.98% probability of landing on heads.

Over many runs, the sequence of measurement outcomes should have accumulated enough uncertainty that truly random bits could be extracted through clever post-processing. However, no existing methods for analysing such sequences would have been efficient enough to reach this goal. Bierhorst et al. therefore introduced a powerful statistical technique, tailored to the weak Bell-inequality violations they observed, that achieved this aim. Ultimately, the authors were able to generate 1,024 random bits in about 10 minutes of data acquisition — corresponding to the measurement of 55 million photon pairs.

Bierhorst and co-workers' random-number generator represents the most meticulous and secure method for producing randomness that has ever been demonstrated. However, its generation rate is much lower than in more-conventional commercial quantum random-number generators, which can produce millions of random bits per second[12]. Nevertheless, improvements in the generation rate can reasonably be expected to the point at which this will no longer be a strong limiting factor.

More problematic is the size of the authors' random-number generator: it is comprised

of measurement stations that are 187 metres apart to prevent subluminal signalling between the photon pairs. This distance might be reduced in the future, but it is hard to imagine how it could reach the dimensions of more-standard electronic hardware (at most, a few centimetres) using foreseeable technology.

Although Bierhorst and colleagues' study will therefore not directly lead to practical, consumer-grade random-number generators, it sets a new direction and ideal for the secure production of random bits. The authors' approach and theoretical methods could be adapted to much more practical and simple designs for random-number generators that potentially retain many of the conceptual and security benefits of their work. ■

Stefano Pironio *is in the Quantum Information Laboratory, Université libre de Bruxelles, 1050 Brussels, Belgium.*
e-mail: stefano.pironio@ulb.ac.be

1. Heninger, N., Durumeric, Z., Wustrow, E. & Halderman, J. A. *Proc. 21st USENIX Security Symp.* 205–220 (USENIX, 2012).
2. Bierhorst, P. *et al. Nature* **556**, 223–226 (2018).
3. Becker, G. T., Regazzoni, F., Paar, C. & Burleson, W. P. in *Cryptographic Hardware and Embedded Systems – CHES 2013* 197–214 (Springer, 2013).
4. Shannon, C. E. *Bell Syst. Tech. J.* **28**, 656–715 (1949).
5. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. *Rev. Mod. Phys.* **86**, 419–478 (2014).
6. Acín, A. & Masanes, L. *Nature* **540**, 213–219 (2016).
7. Pironio, S. *et al. Nature* **464**, 1021–1024 (2010).
8. Liu, Y. *et al. Phys. Rev. Lett.* **120**, 010503 (2018).
9. Hensen, B. *et al. Nature* **526**, 682–686 (2015).
10. Giustina, M. *et al. Phys. Rev. Lett.* **115**, 250401 (2015).
11. Shalm, L. K. *et al. Phys. Rev. Lett.* **115**, 250402 (2015).
12. Herrero-Collantes, M. & Garcia-Escartin, J. C. *Rev. Mod. Phys.* **89**, 015004 (2017).

OPTICAL PHYSICS

Mirrors made of a single atomic layer

Researchers have demonstrated that atomically thin materials can be highly reflective, contrary to general thinking. This finding could have technological implications for nanophotonics, optoelectronics and quantum optics.

KIN FAI MAK & JIE SHAN

The discovery of a single layer of carbon atoms, known as graphene[1], led to great interest in 2D materials. Whereas graphene is highly transparent to visible light[2], 2D materials that are highly reflective could be used as lightweight mirrors in optical or optoelectronic systems. The existence of such materials has been questioned, but, writing in *Physical Review Letters*, Back *et al.*[3] and Scuri *et al.*[4] report that single layers of molybdenum diselenide can have high levels of reflectance.

The importance of the authors' work can be understood by considering the reflection of light from a homogeneous, free-standing thin film of material. When a wave of light of a particular colour — or, equivalently, frequency — hits the film, the oscillating electric field that is associated with the light wiggles the charged particles in the material. This drives the oscillation of electric dipoles (separations between positively and negatively charged particles) at the same frequency as that of the incident light (Fig. 1a).

The oscillating dipoles re-radiate light waves in both the forward and backward directions with respect to the direction of the incident wave. Whereas the latter occurrence gives rise to reflection, the former destructively

interferes with the incident wave, producing transmitted light that has a lower intensity than that of the incident light. The material's response to an oscillating electric field is, in general, not uniform with respect to incident waves from across the electromagnetic spectrum. At a particular frequency, the dipoles have a large oscillation amplitude — a phenomenon known as resonance — which results in more reflection and less transmission of light than at any other frequency.

Like all oscillators in real physical systems, the oscillations of the dipoles are damped, which means that they die out if the event that drives them is stopped. There are two ways in which the energy that is stored in the dipoles can be lost: it can be re-radiated (as discussed previously) or it can be absorbed by the material and converted into heat. These processes are known as radiative and non-radiative damping, respectively. In most materials, both mechanisms of damping operate. The incident light is therefore partly reflected, partly absorbed and partly transmitted.

However, in a material in which radiative damping dominates, the absorption losses would be negligible, and all of the incident electromagnetic energy would be re-radiated. Furthermore, the re-radiation in the forward direction would perfectly cancel out the incident light, through destructive
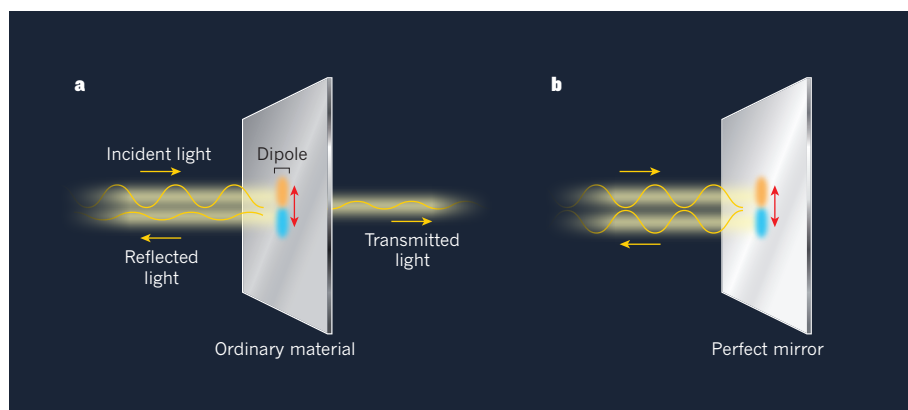
Figure 1 | A conventional material versus a perfect mirror. a, When a wave of light hits a thin film of an ordinary material, it produces electric dipoles — separations between positively charged (orange) and negatively charged (blue) particles. These dipoles oscillate (red arrow) at the same frequency as that of the incident light. They re-radiate light waves in both the forward and backward directions, which gives rise to transmission and reflection, respectively. **b,** By contrast, in a hypothetical perfect mirror, there is no absorption or transmission, and the incident light is reflected entirely. Back et al.[3] and Scuri et al.[4] report near-perfect mirrors made of a single layer of the material molybdenum diselenide.

interference. Owing to conservation of energy, the incident light would be reflected entirely, and the material would act as a perfect mirror (Fig. 1b). This holds true even when the material comprises a single layer of atoms, provided that the oscillating dipoles are being driven at their resonance frequency.

Although theoretical studies have suggested that such conditions could be realized in a 2D array of ultracold atoms[5,6], the authors demonstrate near-perfect mirrors in a solid-state system. They use a single layer of molybdenum diselenide, which is a semiconductor and belongs to a family of materials known as the transition-metal dichalcogenides. In such materials, the oscillating dipoles generated by the incident light are excitons[7] — bound pairs of an electron and a hole (the absence of an electron). The more tightly bound the excitons are, the larger the radiative damping will be, and the more perfectly the mirror will behave. Previous experimental work has shown that the exciton binding in single-layer transition-metal dichalcogenides is extremely strong[8–10], which results in a rate of radiative damping that is much greater than that of conventional semiconductors.

Back et al. and Scuri et al. fabricated high-quality samples of single-layered molybdenum diselenide by encapsulating the material in atomically thin, inert films of hexagonal boron nitride, and then carried out their experiments at a low temperature (4 kelvin). Under these conditions, the authors show that radiative damping of the excitons is the dominant process. They demonstrate mirrors that can reflect a considerable proportion of the incident light — up to 85% in Scuri and colleagues' study — at the exciton resonance frequency of the material.

Although the authors' near-perfect mirrors work only in light from a narrow range of the electromagnetic spectrum (in the vicinity of the resonance frequency), the two studies

open up intriguing possibilities for the fields of nanophotonics and quantum optics. For instance, quantum nonlinear optics requires strong interactions between photons at the single-photon level, which is difficult to achieve in conventional materials. The authors' work shows that quantum nonlinear optics could be realized in single-layer transition-metal dichalcogenides because of the extremely strong light–matter interactions that can be achieved[11].

The authors also demonstrate that the application of a voltage causes the mirrors to switch from being highly reflective to highly transparent. Such mirrors could therefore be used as light modulators, or as other

reconfigurable components, in optical and optoelectronic systems. Moreover, the excitons in single-layer transition-metal dichalcogenides have a feature known as the internal-valley degree of freedom[7], which might enable the mirrors' reflectance to be controlled purely by varying the polarization of the incident light.

About a decade ago, during the early stages of research on 2D materials, many scientists were asking whether a single layer of atoms could be highly reflective. Thanks to Back et al. and Scuri et al., we now know that the answer is yes. ■

Kin Fai Mak and **Jie Shan** are in the Laboratory of Atomic and Solid State Physics and the School of Applied and Engineering Physics, Cornell University, Ithaca, New York 14853, USA.
e-mails: kinfai.mak@cornell.edu; jie.shan@cornell.edu

1. Novoselov, K. S. et al. Science **306**, 666–669 (2004).
2. Casiraghi, C. et al. Nano Lett. **7**, 2711–2717 (2007).
3. Back, P., Zeytinoglu, S., Ijaz, A., Kroner, M. & Imamoğlu, A. Phys. Rev. Lett. **120**, 037401 (2018).
4. Scuri, G. et al. Phys. Rev. Lett. **120**, 037402 (2018).
5. Bettles, R. J., Gardiner, S. A. & Adams, C. S. Phys. Rev. Lett. **116**, 103602 (2016).
6. Shahmoon, E., Wild, D. S., Lukin, M. D. & Yelin, S. F. Phys. Rev. Lett. **118**, 113601 (2017).
7. Mak, K. F. & Shan, J. Nature Photon. **10**, 216–226 (2016).
8. He, K. et al. Phys. Rev. Lett. **113**, 026803 (2014).
9. Chernikov, A. et al. Phys. Rev. Lett. **113**, 076802 (2014).
10. Ye, Z. et al. Nature **513**, 214–218 (2014).
11. Zeytinoğlu, S., Roth, C., Huber, S. & İmamoğlu, A. Phys. Rev. A **96**, 031801 (2017).

This article was published online on 9 April 2018.

PLANT BIOLOGY

Peptide signal alerts plants to drought

It is thought that plants sense water availability in the soil as a way of anticipating drought. The identification of a peptide expressed when water is scarce offers a chance to unravel the underlying molecular mechanism. SEE LETTER P.235

ALEXANDER CHRISTMANN & ERWIN GRILL

Because plants cannot move to escape unfavourable conditions, they must continuously monitor environmental cues to survive when conditions change. Plants can sense interactions with other organisms, such as bacteria, and can monitor light conditions across the spectrum, from ultraviolet to far red. The molecular mechanisms that facilitate those capacities are well understood. But how plants sense drought, cold and salt has

remained an enigma[1]. On page 235, Takahashi et al.[2] report the identification of a peptide that is generated in response to a water deficit in plants.

Drought, cold or salty conditions can affect a plant's water status. Such conditions result in the synthesis[3] of the hormone abscisic acid (ABA), which can regulate the plant's water levels. Stomatal pores in leaves enable plants to take up the carbon dioxide required for photosynthesis, but water vapour can escape through them. ABA can trigger a reduction in

how fully stomatal pores are opened[4], helping to conserve water.

The molecular basis of the link between water deficit and the induction of ABA synthesis has been a mystery. Using the plant *Arabidopsis thaliana* as a model system, Takahashi and colleagues investigated whether members of the CLE family of secreted peptides might have a role in this process. There are more than 30 members of this family, and they are generated by an enzyme-mediated cleavage event. These peptides are involved in diverse biological processes[5]. For example, CLAVATA3 controls the fate of stem cells, and TDIF regulates the formation of the vasculature, the water-transport tissues of plants[6].

Takahashi *et al.* tested 27 CLE peptides for their ability to stimulate ABA synthesis, which is known[7] to occur in the vasculature in response to drought. Enzymes called NCEDs cleave a carotenoid precursor molecule in the pathway that gives rise to ABA, and the expression of the gene *NCED3* is induced by drought[3]. The authors administered CLE peptides to the roots of plants, and monitored whether this treatment induced *NCED3* in leaves. They found that, at low levels of peptide application, only CLE25 was active in regulating *NCED3* expression. CLE25 treatment resulted in an increase in ABA levels and a decrease in stomatal opening. The *CLE25* gene was rapidly expressed in roots in response to drought, and CLE25-deficient mutant plants failed to induce *NCED3* expression in response to dehydration. The formation of CLE25 in the root or shoot was enough to induce *NCED3* in response to dehydration.

The authors tested groups of receptors known to recognize CLE peptides, and identified the receptor proteins BAM1 and BAM3 as being necessary for CLE25-induced responses. A series of grafting experiments clarified how this system works. If roots containing mutations in both the *BAM1* and *BAM3* genes were grafted to wild-type shoots, the application of CLE25 to the plant's roots led to *NCED3* expression in the shoot. However, if the plant was a graft between wild-type roots and shoots that had mutations in both the *BAM1* and *BAM3* genes, *NCED3* was not expressed in the shoot in response to root application of CLE25.

These results are consistent with a model in which CLE25 expressed in the roots can travel to the leaves and bind to BAM1 or BAM3 (Fig. 1). The authors confirmed this pattern of CLE25 mobility by using a mass-spectrometry technique to identify CLE25 peptides that had travelled from the root to the leaf. Little is known about how CLE peptides travel within the plant, and not all such peptides travel as far as CLE25: CLAVATA3 moves only a few layers of cells[8], for example.

Takahashi and colleagues' findings open up potential avenues for determining the long-sought molecular events that occur when a water deficit is initially sensed. The steps leading to *CLE25* expression in response to dehydration are unknown, and their discovery
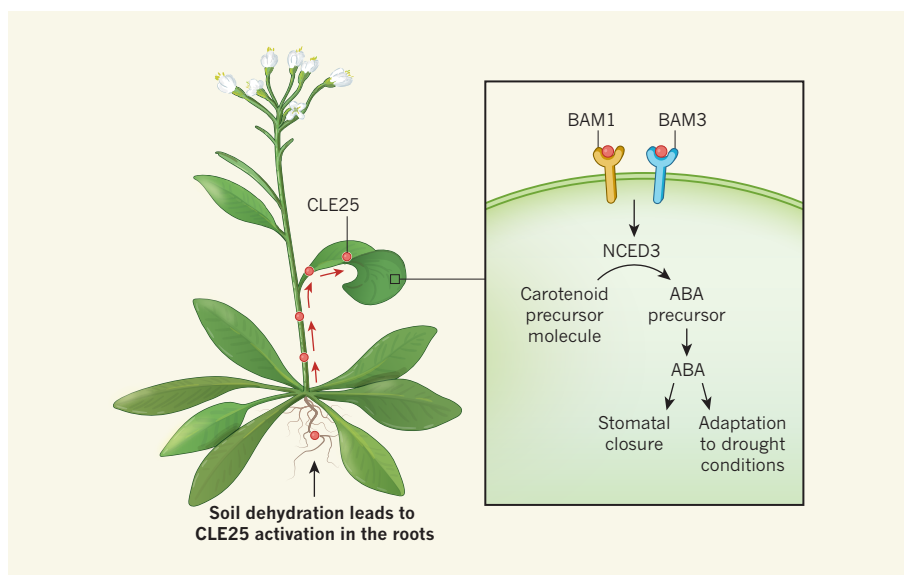
Figure 1 | A peptide aids a plant's response to drought. In drought[3], plants generate the hormone ABA, which can help to regulate plant water levels using processes such as the closure of stomatal pores[4], through which water escapes from leaves. However, the steps that occur between a plant sensing drought and the production of ABA were previously unknown. Using the plant *Arabidopsis thaliana,* Takahashi *et al.*[2] report that the peptide CLE25 is activated in response to drought and is a mobile signal, moving from the roots to the leaves. The authors propose BAM1 and BAM3 proteins as receptors for CLE25, and their results indicate that interactions of CLE25 with these receptors leads to expression of the carotenoid-cleaving enzyme NCED3. The action of this enzyme generates an ABA precursor molecule[3], which is converted to the active ABA signal, facilitating changes that help the plant to cope with a water shortage.

would shed light on this matter. And many questions remain about how CLE25 levels are regulated. How does the presumed cleavage of CLE25 occur? Chemical modifications to CLE25, including the hydroxylation of proline amino-acid residues and possibly the addition of sugar groups, might be crucial for its activity[5]. Whether such modifications are necessary for the function of CLE25 in the drought-sensing process should be investigated.

The molecular mechanism of CLE25 action might be evolutionarily conserved in other plants. The results of Takahashi and colleagues suggest that the CLE25 peptide is generated by the enzymatic cleavage of a precursor protein that generates a 12-amino-acid peptide. In our own analysis of gene sequences, we note that the sequence of this CLE25 peptide in *A. thaliana* is identical to that of many other species, including beet (*Beta vulgaris*), poplar (*Populus trichocarpa*), rice (*Oryza sativa*) and maize (corn; *Zea mays*).

Previous analysis[9] revealed that water deficit can result in tension in the vasculature that can serve as a signal for ABA induction. Transport of CLE25 from the roots to the leaves is likely to be much slower than the immediate relay of this tension cue. Whether this cue and CLE25 act together or independently needs to be addressed. BAM1 and BAM3 are linked to the maintenance of meristem structures, which contain stem cells, and to vasculature development[10]. Whether these receptors use the same signalling pathways for those developmental processes as the ones used in this

drought response also awaits further analysis.

The authors' identification of this role for CLE25 provides an intriguing insight into the regulatory interaction network that plants use to optimize their performance and viability under drought conditions. Water deficit is the major limiting factor for crop yields, and an improved understanding of the molecular strategies used by plants to cope with this environmental challenge[11] might reveal ways of boosting crop resilience and ensuring stability in the future. ∎

Alexander Christmann *and* **Erwin Grill** *are in the Department of Plant Sciences, Technical University of Munich, 5354 Freising, Germany. e-mails: christma@wzw.tum.de; erwin.grill@wzw.tum.de*

1. Zhu, J.-K. *Cell* **167**, 313–324 (2016).
2. Takahashi, F. *et al. Nature* **556**, 235–238 (2018).
3. Nambara, E. & Marion-Poll, A. *Annu. Rev. Plant Biol.* **56**, 165–185 (2005).
4. Kim, T.-H., Böhmer, M., Hu, H., Nishimura, N. & Schroeder, J. I. *Annu. Rev. Plant Biol.* **61**, 561–591 (2010).
5. Matsubayashi, Y. *Annu. Rev. Plant Biol.* **65**, 385–413 (2014).
6. Ito, Y. *et al. Science* **313**, 842–845 (2006).
7. Endo, A. *et al. Plant Physiol.* **147**, 1984–1993 (2008).
8. Lenhard, M. & Laux, T. *Development* **130**, 3163–3173 (2003).
9. Christmann, A., Grill, E. & Huang, J. *Curr. Opin. Plant Biol.* **16**, 293–300 (2013).
10. DeYoung, B. J. *et al. Plant J.* **45**, 1–16 (2006).
11. Yang, Z. *et al. Proc. Natl Acad. Sci. USA* **113**, 6791–6796 (2016).

OCEAN SCIENCE

North Atlantic circulation slows down

Evidence suggests that the circulation system of the North Atlantic Ocean is in a weakened state that is unprecedented in the past 1,600 years, but questions remain as to when exactly the decline commenced. SEE ARTICLE P.191 & LETTER P.227

SUMMER K. PRAETORIUS

The warm, salty waters of the Gulf Stream make a northeasterly meander across the Atlantic Ocean, eventually forming the North Atlantic Current, which then funnels into the Nordic Seas. In the chill of winter, these waters cool and descend with the heavy load of their salinity. This deep convection is a key part of the Atlantic meridional overturning circulation (AMOC; Fig. 1), which can be thought of as an ocean conveyor belt that releases heat to the atmosphere above the North Atlantic Ocean before travelling through the abyssal ocean to resurface in other areas of the world[1].

Given the importance of the AMOC to heat exchange between the ocean and the atmosphere, the varying strength of this system is thought to have major impacts on the global climate, and has been implicated widely in some of the most remarkable and abrupt climate changes of the past[2]. Direct measurements of the modern AMOC flow rates show a decline in its strength in the past decade[3]. Reconstructions of the natural variability and long-term trends of the AMOC are needed, however, to put these recent changes in context. In this issue, Caesar et al.[4] (page 191) and Thornalley et al.[5] (page 227) report on past AMOC variability using different approaches. Both conclude that the modern AMOC is in an unusually subdued state, but they diverge in the details of how and when the AMOC's decline commenced.

Caesar and colleagues inferred changes in the strength of the AMOC in the past century from patterns of anomalies in sea surface temperature (SST) that arise in the North Atlantic when the AMOC weakens. The weakening leads to a warming in the Gulf Stream and a cooling in the subpolar gyre — the cyclonic system of wind-driven ocean currents that lies to the south of Iceland (Fig. 1). Although the link between the relatively cool SSTs of the North Atlantic's subpolar gyre and a slowdown of the AMOC have been studied previously[6–8], the main advance of Caesar and colleagues' work is their comprehensive comparison of global SST data sets with state-of-the-art, high-resolution climate models.

The authors' data analysis shows that this bipolar pattern of cooling and warming emerged in the mid-twentieth century. When they performed climate simulations under a 1% yearly increase in carbon dioxide, the model produced a pattern of SST anomalies in the North Atlantic similar to that seen in

Figure 1 | The Atlantic meridional overturning circulation (AMOC) and the subpolar gyre. The AMOC is an ocean circulation system that consists of warm surface currents (orange) and cold deep-water return flows (blue), as shown in this simplified representation. The surface currents include the Gulf Stream, which feeds a branch of the AMOC known as the North Atlantic Current. The deep-water return flows start from three branches that merge into the North Atlantic Deep Water. Thornalley et al.[5] used measurements of silt in sediment cores to reconstruct the flow speed of the AMOC in the past 1,600 years; the black star indicates the approximate location at which the sediment cores were collected. Caesar et al.[4] analysed temperature anomalies in the North Atlantic subpolar gyre (dashed line) to infer changes in AMOC flow in the past century. Both studies conclude that the AMOC has weakened by about 15% during the periods considered, but they differ on when the flow started to decline.

the observational data, and demonstrated that this pattern was associated with a decline in AMOC strength. The authors then calibrated the model's results with their SST data to estimate that the AMOC has declined by about 15% in the past half-century. They infer that the slowdown in the AMOC was probably a response to warming caused by anthropogenic greenhouse-gas emissions. A possible mechanism could be enhanced melting of the Greenland Ice Sheet[7], which adds fresh water to the surface ocean and reduces the density of the water that drives deep convection.

Thornalley et al. provide a longer-term perspective on changes in AMOC strength during the past 1,600 years using a proxy measurement — the 'sortable-silt' grain size[9] — of deep-sea sediment cores that reflects the speeds of the bottom waters that flow along the path of the North Atlantic Deep Water, the deep-water return flow of the AMOC (Fig. 1). They combined this approach with a method similar to that used by Caesar and colleagues: they used past, near-surface temperature anomalies recorded in the marine sediments to provide additional constraints on the AMOC.

The researchers found that the strength of the AMOC was relatively stable from about AD 400 to 1850, but then weakened around the start of the industrial era. This transition coincides with the end of the Little Ice Age — a multicentennial cold spell that affected many regions of the globe[10]. Thornalley and colleagues infer that the weakening of the AMOC at that time was probably a result of the input of fresh water from the melting of Little Ice Age glaciers and sea ice. They estimate that the AMOC declined in strength by about 15% during the industrial era, relative to its flow in the preceding 1,500 years. This is remarkably similar to Caesar and co-workers' estimate, despite the different time periods on which they base their estimates.

However, the roughly 100-year difference in the proposed timing of the start of the AMOC decline in these two studies has big implications for the inferred trigger of the slowdown. Caesar et al. clearly put the onus on anthropogenic forcing, whereas Thornalley et al. suggest that an earlier decline in response to natural climate variability was perhaps sustained or enhanced through further ice melting associated with anthropogenic global warming. Nevertheless, the main culprit in both scenarios is surface-water freshening.

The two studies are classic examples of 'top-down' and 'bottom-up' approaches, and so it is unsurprising that there is some misalignment
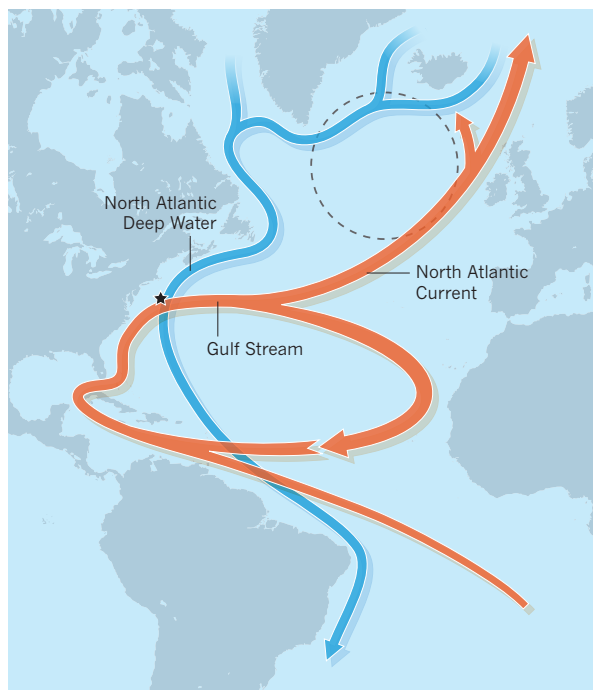
between them. Caesar *et al.* take the top-down approach: their inferences of changes in the AMOC strength are made from reconstructions of regional and global SSTs that are derived from direct measurements of temperature. It is possible that regions other than the North Atlantic in which there has been decadal-scale variability in SSTs could influence the mean global SST from which the AMOC strength is calculated — although the authors do attempt to quell such doubts by showing that the subpolar-gyre SST anomaly is robust relative to the global mean SST for a subset of time periods (see Extended Data Fig. 2 in ref. 4).

Thornalley and colleagues' strategy is more of a bottom-up approach: they use a proxy for deep-water current strength to measure AMOC strength more directly than do Caesar and co-workers. The weaknesses of this approach are that it accounts for only the local bottom currents at the sites from which the cores are taken, which might not capture the entire AMOC system, and that it could be susceptible to local nonlinear effects such as abrupt shifts in the position of the current. However, Thornalley *et al.* show that there is a striking correlation between their

grain-size proxy and the measured density of the Labrador Sea Water (a major component of the North Atlantic Deep Water), as well as with the heat content of the subpolar gyre; these correlations shore up the bridge that links their localized proxy measurements to broader-scale changes in the AMOC.

For now, the timing of the AMOC decline remains a source of intrigue. Future studies that provide a more-detailed spatial network of multiple proxy reconstructions will help to clarify some of the remaining ambiguity about which elements of the AMOC were changing and when. It is — at least scientifically — reassuring to see that the present two studies converge on the conclusion that the modern AMOC is in a relatively weak state. However, in the context of future climate-change scenarios and a possible collapse in the AMOC[11] in response to the continued melting of the Greenland Ice Sheet[12], it is perhaps less reassuring, because a weakened

"The two estimates of the decline are remarkably similar, despite the different time periods on which they are based."

AMOC might lead to considerable changes in climate and precipitation patterns throughout the Northern Hemisphere[13]. ■

Summer K. Praetorius *is in the Geology, Minerals, Energy, and Geophysics Science Center, US Geological Survey, Menlo Park, California 94025, USA.*
e-mail: spraetorius@usgs.gov

1. Broecker, W. S. *Oceanography* **4**, 79–89 (1991).
2. Henry, L. G. *et al. Science* **353**, 470–474 (2016).
3. Srokosz, M. A. & Bryden, H. L. *Science* **348**, 1255575 (2015).
4. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. *Nature* **556**, 191–196 (2018).
5. Thornalley, D. J. R. *et al. Nature* **556**, 227–230 (2018).
6. Drijfhout, S., van Oldenborgh, G. J. & Cimatoribus, A. *J. Clim.* **25**, 8373–8379 (2012).
7. Rahmstorf, S. *et al. Nature Clim. Change* **5**, 475–480 (2015).
8. Menary, M. B. & Wood, R. A. *Clim. Dyn.* **50**, 3063–3080 (2017).
9. McCave, I. N., Manighetti, B. & Robinson, S. G. *Paleoceanography* **10**, 593–610 (1995).
10. PAGES 2k Consortium. *Nature Geosci.* **6**, 339–346 (2013).
11. Liu, W., Xie, S.-P., Liu, Z. & Zhu, J. *Sci. Adv.* **3**, e1601666 (2017).
12. Bakker, P. *et al. Geophys. Res. Lett.* **43**, 12252–12260 (2016).
13. Jackson, L. C. *et al. Clim. Dyn.* **45**, 3299–3316 (2015).

The telomerase enzyme and liver renewal

Cell–tracing analysis reveals that a disperse group of cells in the mouse liver express the enzyme telomerase, which preserves chromosome ends. These cells contribute to liver maintenance and regeneration. SEE LETTER P.244

KENNETH S. ZARET

The enzyme telomerase maintains the length of specialized repetitive structures called telomeres, which are found at the ends of chromosomes. When they become damaged or shortened, telomeres can stop cells from dividing[1]. Most cells in adult humans have very low or undetectable levels of telomerase and relatively short telomeres, and therefore have a limited ability to replicate[2]. However, elevated telomerase levels are seen in various animal and human stem cells that must retain their replicative capacity for self-renewal[3]. Telomerase defects are associated with tissue scarring (fibrosis) in the livers of both mice and humans[4,5], but which cells in the liver express telomerase, and whether they act as stem cells, has been unclear. On page 244, Lin *et al.*[6] characterize this cell population in mice.

First, the authors identified telomerase-expressing cells in the mouse liver and tracked

descendent cells. The group genetically engineered mice to contain a modified version of the gene *Tert*, which encodes a subunit of telomerase. When the mice are treated with a drug, this alteration causes cells expressing *Tert* to be indelibly labelled by a fluorescent protein. Once the genetically modified cells are triggered in this way, they and all their descendants produce the fluorescent protein, even if the cells no longer express *Tert* itself.

Lin *et al.* found that 3–5% of hepatocytes, the most prevalent type of cell in the liver, fluoresce in response to drug treatment. The authors confirmed, by quantitation of messenger RNA levels, that these cells express *Tert*. Next, they examined the livers of adult mice one year after drug treatment. The initially labelled cells (dubbed Tert^High) had given rise to clusters of descendants dispersed throughout the liver's lobes, making up about 30% of the liver's total mass (Fig. 1). Adult hepatocytes die and are replaced infrequently, so the increase in labelled cells over long periods indicates

that the Tert^High hepatocytes contribute to the gradual renewal of the liver under normal conditions.

A key question is whether the Tert^High hepatocytes are a stable, self-renewing population. Alternatively, *Tert* could be expressed in certain cells for a period of time, then shut off in those hepatocytes and expressed in others. In support of the former case, when Lin *et al.* triggered fluorescent-protein labelling three times over a ten-week period, they found that the numbers of labelled hepatocytes were comparable to those for a single trigger. Next, they showed that 75% of labelled hepatocytes expressed high levels of *Tert* mRNA when they were examined a month after a single drug treatment, whereas only 18% did so after a year, indicating that, as the population gradually expands, Tert^High cells not only self-renew but also give rise to progeny that do not express *Tert* (Tert^Low). Finally, the researchers demonstrated that Tert^High hepatocytes proliferate more than Tert^Low cells, whereas Tert^Low cells exhibit higher expression of genes relating to metabolism and biosynthesis than do Tert^High cells.

Taking these data together, the authors suggest that Tert^High hepatocytes behave like stem cells. But before concluding that the Tert^High cells are bona fide stem cells for the liver, it will be necessary to determine whether the Tert^High population becomes exhausted or remains at similar levels in older mice (because hepatocytes are still renewed in ageing mice), and whether Tert^Low cells convert to Tert^High over longer periods than those used here (which would indicate that this population is not acting as stem cells). It will also be interesting
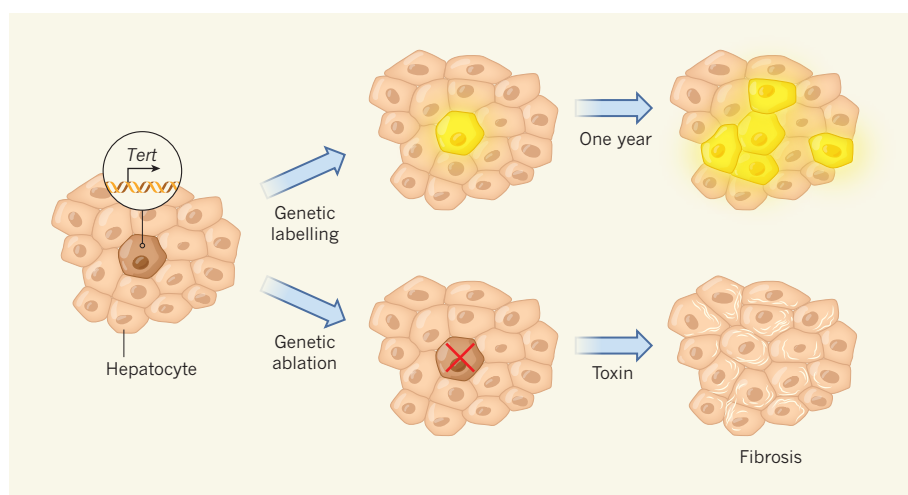
Figure 1 | Lineage tracing in the liver. Lin *et al.*[6] characterize the hepatocyte cells in the mouse liver that express high levels of the gene *Tert*, which encodes a subunit of the enzyme telomerase. The authors generated mice that carry a genetically engineered version of *Tert*: when the mice are treated with a drug, any cells expressing *Tert* are indelibly labelled with a fluorescent protein. Those cells and all their descendants fluoresce, and so can be tracked. Only 3–5% of cells fluoresced immediately after drug treatment. One year later, about 30% of cells fluoresced, but most of these did not express *Tert*, indicating that the rare *Tert*-expressing cells give rise to new hepatocytes to help regenerate the liver. If the *Tert*-expressing cells are genetically ablated, the liver is susceptible to scarring (fibrosis) after toxin damage.

to determine the processes by which cells transition from Tert^High to Tert^Low, and how this change relates to homeostatic control of liver mass.

Importantly, stem cells typically reside in a special tissue compartment, or niche, that supports their regenerative capacity. Yet the Tert^High cells are dispersed throughout the liver. This dispersal of Tert^High cells is interesting because hepatocytes reside in different zones in each lobe of the liver, and earlier studies[7] implicated one zone or another as being more relevant to liver regeneration. By contrast, Lin *et al.* provide evidence for a 'distributed model' for hepatocyte renewal. The research indicates that, although the Tert^High hepatocytes possess features of stem cells, those features are not of a conventional type.

In the past three years, one regenerative hepatocyte population near the central vein has attracted particular attention. The population responds to venous signals to self-renew during homeostasis, producing progeny that migrate outwards from the central zone[8]. Lin *et al.* found a few Tert^High hepatocytes in the central zone in healthy livers, but these cells did not reside close enough to the central vein to respond to its signals. However, when the authors damaged the central-vein zone, Tert^High descendants appeared there and responded to venous signals. Moreover, after damage to the liver tissue in another region, around the portal vein, hepatocytes descended from Tert^High cells appeared abundantly in the periportal and mid-lobular zones, and the researchers found that ablation of Tert^High hepatocytes impaired this regenerative response, leading to liver fibrosis. Taking the above findings together with those of other studies of liver injury, it seems that various types of hepatocyte (as well as cells

from the bile duct)[9–12] can regenerate the mouse liver under a range of damage conditions.

In the future, it will be crucial to assess how relevant these findings in mice are to human liver regeneration. The fact that ablation of Tert^High hepatocytes results in fibrosis in the injured mouse liver seems to support relevance for humans, because people who harbour mutations in *TERT* and genes that encode other telomere-related factors can also exhibit fibrosis and cirrhosis (the latter being a predictor of liver cancer)[5]. However, Tert^High hepatocytes have not been seen in human

livers — although the possibility has not yet been assessed with the sensitivity of the genetic-labelling approach used in mice by Lin and colleagues. An alternative explanation for diseases in humans who have telomerase-related mutations is that excessive telomere shortening in early development might affect many organ progenitors in a nonspecific way.

More-detailed studies in humans will be needed to confirm how telomerase-based regeneration forestalls liver disease, and possibly liver cancer. Nevertheless, Lin and colleagues' study provides insight into a previously unidentified, dispersed-cell mode of liver regeneration. ■

Kenneth S. Zaret *is in the Department of Cell and Developmental Biology, Institute for Regenerative Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.*
e-mail: zaret@pennmedicine.upenn.edu

1. Palm, W. & de Lange, T. *Annu. Rev. Genet.* **41**, 301–334 (2008).
2. Blasco, M. A. *Nature Chem. Biol.* **3**, 640–649 (2007).
3. Günes, C. & Rudolph, K. L. *Cell* **152**, 390–393 (2013).
4. Rudolph, K. L., Chang, S., Millard, M., Schreiber-Agus, N. & DePinho, R. A. *Science* **287**, 1253–1258 (2000).
5. Calado, R. T. *et al.* *PLoS ONE* **4**, e7926 (2009).
6. Lin, S. *et al.* *Nature* **556**, 244–248 (2018).
7. Stanger, B. Z. *Annu. Rev. Physiol.* **77**, 179–200 (2015).
8. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. *Nature* **524**, 180–185 (2015).
9. Yanger, K. *et al.* *Cell Stem Cell* **15**, 340–349 (2014).
10. Tarlow, B. D. *et al.* *Cell Stem Cell* **15**, 605–618 (2014).
11. Font-Burgada, J. *et al.* *Cell* **162**, 766–779 (2015).
12. Raven, A. *et al.* *Nature* **547**, 350–354 (2017).

This article was published online on 4 April 2018.

Backbone of RNA viruses uncovered

The evolutionary history of viruses is largely unknown. Large-scale discovery of vertebrate RNA viruses shows that, although viruses often jump between hosts, most have co-evolved with their hosts over millions of years. SEE ARTICLE P.197

MARK ZELLER & KRISTIAN G. ANDERSEN

Many human diseases, from the common cold to deadly haemorrhagic fevers, are caused by RNA viruses. Most of these viruses are thought to have originated from close relatives that infected mammals[1,2], and so the majority of virus-discovery studies have focused on mammals and birds[3]. RNA viruses, however, are probably older than the last common ancestor of life on Earth[4,5]. Detailed genetic

information for RNA viruses from other classes of vertebrate is sorely needed if we are to fully understand long-term virus evolution. On page 197, Shi *et al.*[6] report the discovery of previously unidentified vertebrate RNA viruses from across evolutionary timescales.

The authors analysed the viruses in 186 vertebrate species using an approach called metatranscriptomic sequencing, in which all of the RNA present in a sample is sequenced. The samples were taken from

species of fish, amphibian and reptile — every vertebrate class except mammals and birds. In these samples, Shi and colleagues discovered a total of 214 viruses, dramatically increasing the number of known RNA viruses in each vertebrate class. For example, they identified more than 20 RNA viruses that infect amphibians, whereas just a few had previously been identified[7–9].

The analysis also revealed an astonishing level of biodiversity — the researchers identified previously unknown viruses in almost every RNA-virus family known to infect mammals. These include viruses highly pathogenic to humans, such as influenza virus, arenaviruses and filoviruses, that have not previously been reported in fish or amphibians.

Shi *et al.* used this information to construct phylogenetic trees that describe the evolutionary relationships between viruses. They found that the phylogenies of RNA viruses were broadly comparable to those of the viruses' vertebrate hosts. This shows that RNA viruses followed a similar evolutionary trajectory to vertebrates, and have co-evolved with their hosts over millions of years (Fig. 1). The evolution of vertebrates began more than 500 million years ago — vertebrate life then divided into several classes of fish, followed by the evolution of amphibians that moved on to land (http://www.onezoom.org). The authors' findings indicate that mammalian RNA viruses probably originated from viruses that infected fish, and then followed vertebrates on to land.

However, the researchers also show that some viruses can infect multiple hosts, indicating that, in addition to co-evolution, viruses have made jumps between species. In fact, many virus outbreaks in humans are the result of animal-to-human transmission, as exemplified by the recent Ebola epidemic in West Africa[10]. Most cross-species transmission events result in limited or no onwards transmission (the virus typically continues to circulate only temporarily in the new host species), and the ability of a virus to establish itself depends on a range of factors, including host divergence[11]. Thus, transmission between animals belonging to the same vertebrate class (bats to humans, for example) is more likely than that between animals belonging to different vertebrate classes (such as reptiles to mammals). But Shi and colleagues' phylogenies reveal that viruses regularly jump between vertebrate classes, with successful onwards transmission that can continue for millions of years.

The current study greatly expands our knowledge of vertebrate virus evolution. However, it is not without limitations. First, excluding birds and mammals, there are more than 50,000 vertebrate species. And although the current study is one of the largest of its kind, Shi *et al.* sampled less than 0.5% of these species. Moreover, the authors focused their sampling towards common taxa such as ray-finned fishes, and included relatively
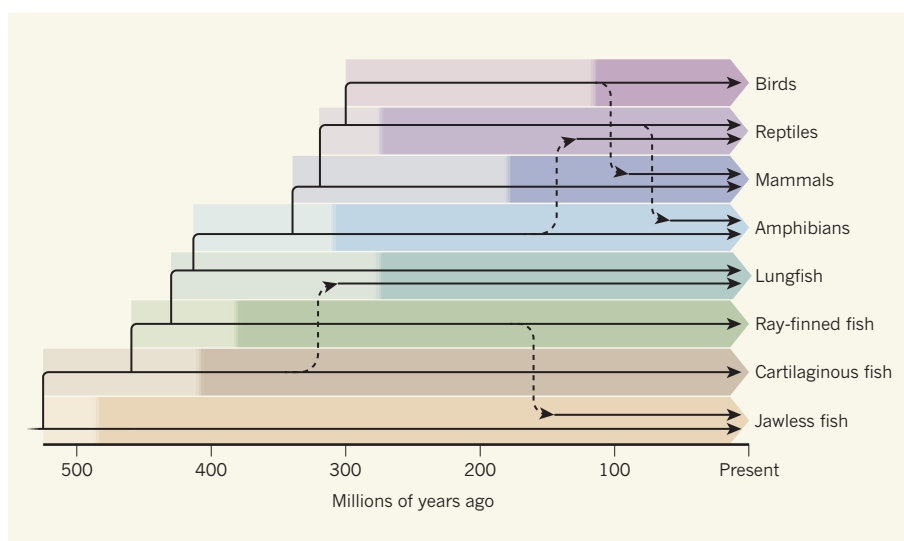
Figure 1 | Tracking the evolution of RNA viruses. Shi *et al.*[6] sequenced RNA viruses present in various classes of vertebrate, and constructed trees of virus evolution. Over a period of 525 million years, vertebrates branched off into several classes. The beginning of each coloured blocked arrow indicates the divergence between a vertebrate group and that below it in the figure; the beginning of the darker shading indicates the time that the most recent common ancestor of currently extant members of a class arose. The authors found that RNA viruses co-diverged with their vertebrate hosts (black lines indicate virus evolution). Each vertebrate class is dominated by its own set of RNA viruses; however, occasional cross-species transmissions occur (dashed arrows), introducing new viruses into a particular class. This phylogenetic tree is a simplified schematic to exemplify RNA-virus evolution as a whole, and does not reflect precise dates or cross-species transmission events found by the authors.

few amphibians. This means that the group's findings represent only a minuscule fraction of the total diversity of RNA viruses. We are just scratching the surface of these viruses' evolutionary history. Our understanding of viral evolution will continue to expand as we sample RNA viruses from across deeper evolutionary timescales.

Another limitation of the current study is that — as is typical for this type of work — new viruses are identified on the basis of genetic similarity to those that have been sequenced previously. This strategy has the potential to introduce biases. It is therefore possible that there are entire groups of viruses yet to be discovered, because they cannot be detected using similarity-based approaches.

Finally, it is becoming increasingly clear that only a tiny fraction of RNA viruses will ever infect humans, and the factors that contribute to virus emergence in humans are not fully understood. As Shi *et al.* show, phylogenetic analyses are a powerful tool for identifying cross-species transmissions that happened in the past. But they cannot be used to predict host jumps and virus emergence of the future — the complexity of successful cross-species transmission renders efforts to predict disease emergence by mapping non-human virus diversity ineffective[12]. Studies that give us a more fundamental understanding of RNA-virus evolution and diversity, as Shi and colleagues work does, will be crucial to inform future surveillance efforts in humans.

It took us many decades to understand the basics of the evolutionary history of

vertebrates. It will probably take even longer before we can confidently say that we are beginning to understand the enormous diversity of RNA viruses and their complex relationships with humans and other vertebrates. Shi *et al.* have provided an exciting starting point from which to strike out towards this goal. ∎

Mark Zeller and **Kristian G. Andersen** *are in the Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, California 92037, USA.* **K.G.A.** *is also in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, and at the Scripps Translational Science Institute. e-mails: zellerm@scripps.edu; andersen@scripps.edu*

1. Wolfe, N. D., Dunavan, C. P. & Diamond, J. *Nature* **447**, 279–283 (2007).
2. Woolhouse, M. E. J. & Brierley, L. *Sci. Data* **5**, 180017 (2018).
3. Olival, K. J. *et al. Nature* **546**, 646–650 (2017).
4. Holmes, E. C. *J. Virol.* **85**, 5247–5251 (2011).
5. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. *Biol. Direct* **1**, 29 (2006).
6. Shi, M. *et al. Nature* **556**, 197–202 (2018).
7. Tristem, M., Herniou, E., Summers, K. & Cook, J. *J. Virol.* **70**, 4864–4870 (1996).
8. Reuter, G. *et al. J. Gen. Virol.* **96**, 2607–2613 (2015).
9. Ip, H. S., Lorch, J. M. & Blehert, D. S. *Emerg. Microbes Infect.* **5**, e97 (2016).
10. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. *Nature* **538**, 193–200 (2016).
11. Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G. & Lemey, P. *Phil. Trans. R. Soc. B* **368**, 20120196 (2013).
12. Geoghegan, J. L. & Holmes, E. C. *Open Biol.* **7**, 170189 (2017).

ARTICLE

Reversible Mn^{2+}/Mn^{4+} double redox in lithium–excess cathode materials

Jinhyuk Lee[1,9]*, Daniil A. Kitchaev[2], Deok–Hwang Kwon[1], Chang–Wook Lee[3], Joseph K. Papp[4], Yi–Sheng Liu[5], Zhengyan Lun[1], Raphaële J. Clément[1], Tan Shi[1], Bryan D. McCloskey[4,6], Jinghua Guo[5,7], Mahalingam Balasubramanian[3] & Gerbrand Ceder[1,8]*

There is an urgent need for low–cost, resource–friendly, high–energy–density cathode materials for lithium–ion batteries to satisfy the rapidly increasing need for electrical energy storage. To replace the nickel and cobalt, which are limited resources and are associated with safety problems, in current lithium–ion batteries, high–capacity cathodes based on manganese would be particularly desirable owing to the low cost and high abundance of the metal, and the intrinsic stability of the Mn^{4+} oxidation state. Here we present a strategy of combining high–valent cations and the partial substitution of fluorine for oxygen in a disordered–rocksalt structure to incorporate the reversible Mn^{2+}/Mn^{4+} double redox couple into lithium–excess cathode materials. The lithium–rich cathodes thus produced have high capacity and energy density. The use of the Mn^{2+}/Mn^{4+} redox reduces oxygen redox activity, thereby stabilizing the materials, and opens up new opportunities for the design of high–performance manganese–rich cathodes for advanced lithium–ion batteries.

Lithium–ion–based energy storage is becoming a pervasive technology in today's society. Introduced in the early 1990s for use in portable electronics, it has now migrated to applications such as transportation and grid, for which energy storage needs will soon dwarf the use in electronics[1,2]. Indeed, today, with electric vehicles making up about 1% of all car sales, almost half of all Li-ion batteries produced are already directed towards transportation. These new applications increase the demand for safe high-energy storage at low cost and without relying on constrained natural resources[1,2]. In this context, it is remarkable that almost all Li-ion cathode materials rely on only two transition metals, Ni and Co, which are the electroactive elements in the layered-rocksalt cathode materials in the $Li(Ni,Mn,Co)O_2$ chemical space (NMCs)[3]. On one end of this compositional spectrum, $LiCoO_2$ dominates the electronics sector, whereas Ni-rich materials are of interest for the automotive sector[2]. Although Mn has been used in a spinel cathode[4], and Fe in the $LiFePO_4$ olivine[5], these compounds suffer from low energy density. Given the limits of energy density that can be achieved with the layered NMCs and the potential resource constraints on cobalt[2], it is of interest to develop high-capacity cathode materials based on other redox metals. In particular, transition metals that can exchange two electrons are of interest for their ability to create high capacity, similar to the Ni^{2+}/Ni^{4+} couple in NMC cathodes. Low cost and low toxicity make the Mn^{2+}/Mn^{4+} couple particularly desirable[6] for designing high-performance Li-ion batteries that are also inexpensive and eco-friendly.

Manganese is currently used in cathode materials, but mostly in the inert Mn^{4+} state, as in NMC cathodes, or for its Mn^{3+}/Mn^{4+} couple, as in $LiMn_2O_4$ spinel[1,3,4,7]. More recently, Mn^{3+} has been used in disordered-rocksalt-type cathodes, such as $Li_{1.3}Mn_{0.4}Nb_{0.3}O_2$, in which the low capacity from Mn^{3+}/Mn^{4+} needs to be overcome by a large amount of oxygen redox[8], which can trigger O loss, resulting in substantial voltage and capacity fade[7,9,10]. In $Li_4Mn_2O_5$, a high initial capacity (>300 mAh g^{-1}) is achieved by oxidizing Mn^{3+} past the standard Mn^{3+}/Mn^{4+} redox couple, but this causes substantial voltage and capacity fade in subsequent cycles[11]. In our approach, we start instead from Mn^{2+} in the discharged state so that a high theoretical capacity can be obtained by oxidizing to Mn^{4+} without relying on O redox. Cycling between two stable valence states of Mn, and limiting the O redox, is expected to improve the reversibility of the charge/discharge process. Reduction to Mn^{2+} has been observed[12] by lithiation of amorphous $Li_{1.5}Na_{0.5}MnO_{2.85}I_{0.12}$, but as this cathode material is synthesized in the charged state it does not enable Li to be brought into the Li-ion cell. The development of a high-performance Li-ion cathode based on the Mn^{2+}/Mn^{4+} couple requires a material that forms in its discharged state, contains enough Mn^{2+} and Li^+ ions to provide high capacity and preferably crystallizes in a dense structure, such as the layered or disordered-rocksalt structure, to maximize its volumetric energy density. Introducing Mn^{2+} in the dense layered or disordered materials has been difficult, as the Li excess ($x > 1$ in $Li_xTM_{2-x}O_2$, where TM is transition metal) required to achieve high practical capacity[10,13,14] demands a high average transition metal valence.

In this work, we demonstrate that high capacity (>300 mAh g^{-1}) and energy density (about 1,000 Wh kg^{-1}) can be achieved in disordered-rocksalt Li-rich intercalation cathodes from Mn^{2+}/Mn^{4+} double redox combined with a small amount of O redox. A critical step is that we are able to lower the Mn valence in the cathode material through a combined strategy of high-valent cation (Nb^{5+}, Ti^{4+}) substitution[8] and O^{2-} replacement[15–17] by F$^-$. This O^{2-} replacement was recently shown to be aided by Li excess and cation disorder[15]. We target the Mn^{2+}-containing compositions $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$, which have a theoretical Mn^{2+}/Mn^{4+} redox capacity of 270 mAh g^{-1} and 230 mAh g^{-1}, respectively. Given the high Mn capacities, only a small amount of O redox is required for these materials to deliver a total capacity over 300 mAh g^{-1}, mitigating problems related to O redox. Thus realized, high capacity from Mn^{2+}/Mn^{4+} double redox

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA, USA. [2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]X-ray Science Division, Advanced Photon Source, Argonne National Laboratory, Argonne, IL, USA. [4]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA. [5]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Energy Storage and Distributed Resources Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA, USA. [8]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [9]Present address: Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. *e-mail: jinhyuk@mit.edu; gceder@berkeley.edu
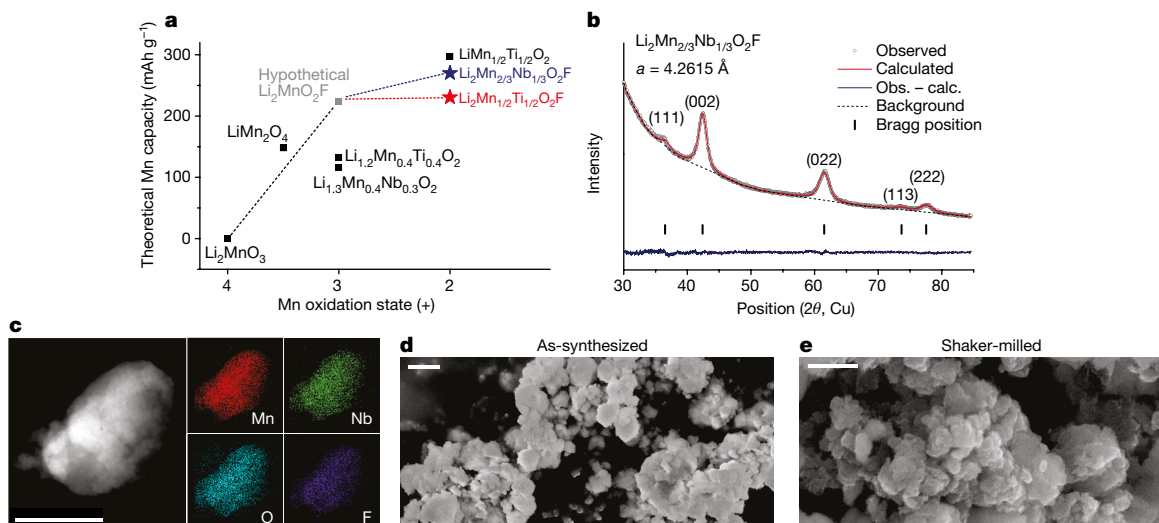
Fig. 1 | Design and structural characterization of $Li_2Mn_{2/3}Nb_{1/3}O_2F$.
a, Theoretical Mn-redox capacity of various Mn-based cathode materials.
b, The X-ray diffraction pattern of $Li_2Mn_{2/3}Nb_{1/3}O_2F$. **c**, EDS mapping
(Mn, Nb, O, F) on a $Li_2Mn_{2/3}Nb_{1/3}O_2F$ particle. Scale bar, 100 nm.
d, **e**, SEM images of $Li_2Mn_{2/3}Nb_{1/3}O_2F$: **d**, as-synthesized (scale bar,
400 nm) and **e**, shaker-milled with carbon black (scale bar, 200 nm).

opens new opportunities for the design of high-performance Li-ion cathode materials.

Structural characterization of $Li_2Mn_{2/3}Nb_{1/3}O_2F$

To evaluate the Mn^{2+}/Mn^{4+} redox strategy, we first test a new disordered Li-rich material: $Li_2Mn_{2/3}Nb_{1/3}O_2F$ (equivalent to $Li_{1.333}Mn_{0.444}Nb_{0.222}O_{1.333}F_{0.667}$), synthesized by a mechanochemical ball-milling method[11,16,17]. The combined presence of high-valent Nb^{5+} and low-valent F^- sets up the charge balance to incorporate Mn as Mn^{2+} in the pristine $Li_2Mn_{2/3}Nb_{1/3}O_2F$ material, leading to a very high theoretical Mn-redox capacity of 270 mAh g^{-1}, which is more than twice that of a typical Mn-based Li-rich cathode material (Fig. 1a). In addition, the d^0 configuration of Nb^{5+} (similar to that of Ti^{4+}, V^{5+}, Zr^{4+} and Mo^{6+}) promotes the formation of a disordered-rocksalt structure[18].

X-ray diffraction (XRD) patterns (Fig. 1b, Extended Data Table 1) and elemental analysis (Extended Data Table 2) show that the compound forms in a disordered-rocksalt phase with a composition close to the target composition[11]. XRD refinement yields a lattice parameter of about 4.262 Å (Extended Data Table 1). Energy-dispersive spectroscopy mapping (EDS) on a $Li_2Mn_{2/3}Nb_{1/3}O_2F$ particle, using a transmission electron microscope (TEM), reveals a uniform distribution of Mn, Nb, O and F (Fig. 1c). Results of ^7Li and ^{19}F nuclear magnetic resonance (NMR) reveal that some Li can be found in diamagnetic environments and some F can be found in LiF-like environments (Extended Data Fig. 1, Methods section 'Supplementary Note 1'). Although this suggests that small amounts of impurity phases (for example LiF, Li_2O, Li_2CO_3) may be present in the as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ sample, we cannot rule out the presence of diamagnetic or LiF-like local domains in the rocksalt phase. In fact, no crystalline impurities could be detected with XRD. TEM shows that the primary particles are polycrystalline and made of crystalline grains about 15 nm in size (Extended Data Fig. 2). No amorphous components were detected in TEM, indicating that the electrochemical properties are predominantly determined by the $Li_2Mn_{2/3}Nb_{1/3}O_2F$ phase. Scanning electron microscopy (SEM) shows that the primary particle size of the as-prepared $Li_2Mn_{2/3}Nb_{1/3}O_2F$ compound is 100–300 nm (Fig. 1d), which is reduced to less than 100 nm after mixing with carbon black using a shaker-mill for electrode fabrication (Fig. 1e).

Electrochemical performance of $Li_2Mn_{2/3}Nb_{1/3}O_2F$

Galvanostatic charge–discharge tests of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ at 20 mA g^{-1} show a discharge capacity of 238 mAh g^{-1} (708 Wh kg^{-1}) between 1.5 V and 4.6 V, which increases to 277 mAh g^{-1} (849 Wh kg^{-1}) and 304 mAh g^{-1} (945 Wh kg^{-1}) with higher charge cut-off voltages of

4.8 V and 5.0 V, respectively (Fig. 2a–c). In a test between 1.5 V and 5.0 V at 10 mA g^{-1} (Fig. 2d), the discharge capacity further increases to 317 mAh g^{-1}, delivering a very high energy content of 995 Wh kg^{-1} (3,761 Wh l^{-1}). This discharge capacity of about 320 mAh g^{-1} and specific energy approaching 1,000 Wh kg^{-1} are among the highest values achieved by Li-ion intercalation cathodes[10–12,19]. The reversible capacity and energy density at 20 mA g^{-1} decrease to 233 mAh g^{-1} (180 mAh g^{-1}) and 760 Wh kg^{-1} (600 Wh kg^{-1}), respectively, as the voltage window is reduced to 2.0–4.8 V (2.3–4.6 V) (Extended Data Fig. 3). The rate capability of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ is fairly good. Figure 2e compares the first cycle profiles of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ under different rates between 1.5 V and 5.0 V. The material delivers as high as 226 mAh g^{-1} (695 Wh kg^{-1}) at 200 mA g^{-1} and up to 140 mAh g^{-1} (410 Wh kg^{-1}) at a very high rate of 1,000 mA g^{-1} (Extended Data Fig. 4). The data presented here were obtained on electrode films made of 60 wt% active material, but the performance is similar for electrodes with a higher loading of 70 wt% and 80 wt% (Extended Data Fig. 3).

The voltage profiles of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ do not contain significant hysteresis and remain nearly identical during cycling, suggesting that structural changes and oxygen loss are small[8,20,21]. Only upon very high-voltage charging to above 4.7 V is an apparent voltage plateau observed which is barely seen in discharge (Fig. 2f). As $Li_2Mn_{2/3}Nb_{1/3}O_2F$ delivers a higher capacity than its theoretical Mn capacity (270 mAh g^{-1}), we expect that the charge plateau at about 4.8 V accompanies O-oxidation. The asymmetry in charge/discharge voltage is similar to previous observations in which the O-oxidation plateau is not recovered in the discharge[8,20,21]. Nevertheless, this plateau in $Li_2Mn_{2/3}Nb_{1/3}O_2F$ appears only after charging above about 250 mAh g^{-1}, leading to less voltage hysteresis than for the other Mn-redox-active disordered compounds in which the O-oxidation plateau occurs typically much earlier in the charge[8,20,21]. The smaller amount of O-oxidation and negligible changes in the voltage profile of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ are further supported by differential electrochemical mass spectrometry (DEMS) results, which show negligible O_2 (g) evolution up to a charge of 5 V (Extended Data Fig. 5, Methods section 'Supplementary Note 2'). In addition, voltage fade is small in this material (Extended Data Fig. 6, Methods section 'Supplementary Note 3'). These results indicate that our strategy of going to Mn^{2+} compounds to increase the Mn-redox capacity at the expense of O redox is successful in increasing capacity and reversibility.

Redox mechanism of $Li_2Mn_{2/3}Nb_{1/3}O_2F$

The redox mechanism and structural evolution of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ have been further studied by X-ray diffraction, and by hard X-ray and
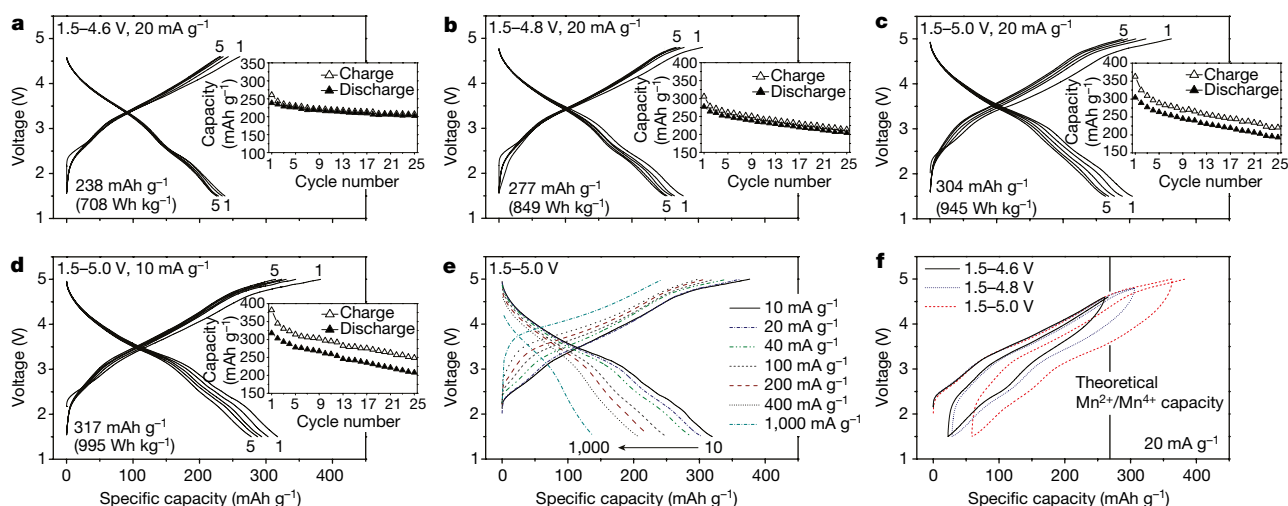
Fig. 2 | Electrochemical performance of Li₂Mn₂/₃Nb₁/₃O₂F. a–d,Voltage profiles and capacity retention of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ under various cycling conditions: **a**, 1.5–4.6 V, 20 mA g⁻¹; **b**, 1.5–4.8 V, 20 mA g⁻¹; **c**, 1.5–5.0 V, 20 mA g⁻¹; and **d**, 1.5–5.0 V, 10 mA g⁻¹. **e**, The first-cycle voltage profiles of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ when cycled between 1.5 V and 5.0 V at 10, 20, 40, 100, 200, 400 and 1,000 mA g⁻¹. **f**, The first-cycle and second-charge profiles of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ under different voltage windows: 1.5–4.6 V, 1.5–4.8 V and 1.5–5.0 V. All tests were conducted at room temperature.

soft X-ray absorption spectroscopies. Figure 3a shows a reversible lattice-parameter change upon cycling, as observed in other disordered-rocksalt intercalation cathodes[8,20–22]. The shift of the (002) and (022) peaks to a higher angle upon charge (indicating a decrease of lattice parameters) is recovered on discharge. Upon charging, the a lattice parameter decreases from 4.258 Å to 4.130 Å at the top of charge and returns to 4.250 Å after full discharge.

Hard X-ray absorption spectroscopy (XAS) suggests that Mn^{2+} is oxidized during charge towards Mn^{4+}, a process that is reversed upon discharge. Figure 3b shows the Mn K-edge X-ray absorption near-edge structure (XANES) for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ at various states of charge and discharge. As the charge capacity increases from 0 to 135 mA h g⁻¹ and 270 mAh g⁻¹, the Mn K-edge shifts from an energy close to MnO (Mn^{2+} reference) to that of Mn_2O_3 (Mn^{3+} reference) and then partway to the energy seen in MnO_2 (Mn^{4+} reference). Further charging to 360 mAh g⁻¹ leads to only minor shifts. The Mn K-edge almost completely returns to its original position after discharge. Although the exact amount of each valence state cannot be quantified, as the near-edge structure is sensitive to both the oxidation state and bonding environment[23], this result suggests that, on full charge, Mn^{2+} is oxidized to Mn^{4+} with some Mn^{2+} or Mn^{3+} ions remaining. Full recovery to Mn^{2+} occurs on discharge. This interpretation is further supported by a derivative analysis on the Mn pre-edge at about 6,540 eV (Extended Data Fig. 7, Methods section

'Supplementary Note 4'). The species Nb^{5+} does not participate in redox processes (Extended Data Fig. 8).

Whereas hard X-rays probe metal oxidation, soft X-ray absorption using a total fluorescence yield can be used to investigate oxygen redox behaviour. Figure 3c traces the pre-edge features of the O K-edge XAS spectra of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ at various states of charge. The pre-edge is primarily associated with the O 1s to 2p transition, and its intensity is attributable to the density of unoccupied Nb 4d–O 2p and Mn 3d–O 2p hybridized states. We associate the pre-edge feature around 530.9 eV with unoccupied Nb 4d–O 2p hybridized states, as Mn^{2+} oxides (for example MnO) typically exhibit a pre-edge feature above[24,25] about 533 eV. Charging to 135 mAh g⁻¹ (theoretical Mn^{2+}/Mn^{3+} limit) increases the intensity in the 529–532 eV range which is typical for Mn^{3+} oxides such as[21,25] Mn_2O_3. After charging to 270 mAh g⁻¹ and 360 mAh g⁻¹, an intensity gain is observed broadly between 528 and 530 eV. The largest major intensity gain is centred around 529 eV (feature A) which is characteristic of Mn^{4+} oxides (for example MnO_2, Li_2MnO_3)[24–26], suggesting some Mn^{3+}/Mn^{4+} oxidation on charge. Along with this feature A, we see an intensity gain at 530–531 eV (feature B) after charging to 270 mAh g⁻¹ and 360 mAh g⁻¹. Previously, O oxidation has been shown to create a broad component around 530.2 eV in Mn-based disordered Li-rich cathodes[8,21]. Therefore, this feature B most probably indicates O oxidation which concurrently occurs with Mn^{3+}/Mn^{4+} oxidation. Discharging to 320 mAh g⁻¹ restores the
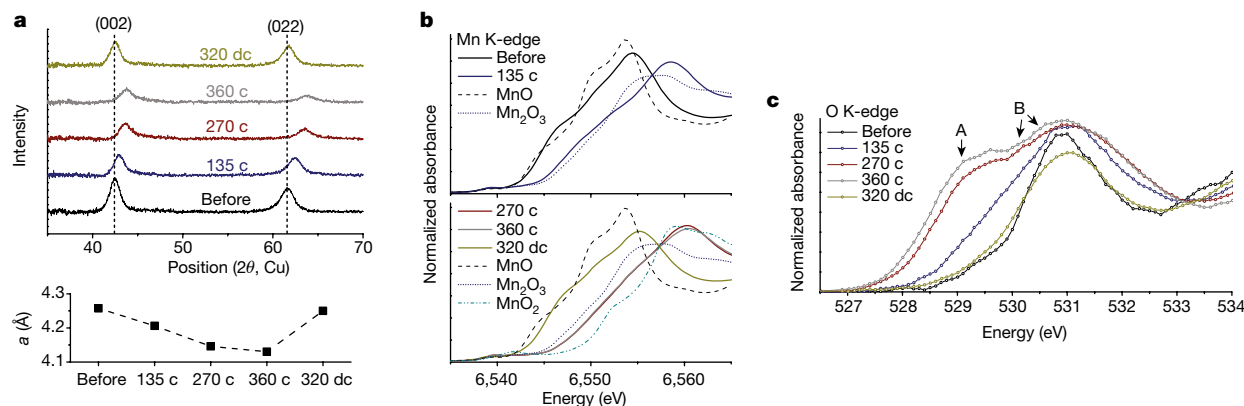
Fig. 3 | Reaction mechanism of Li₂Mn₂/₃Nb₁/₃O₂F. a, XRD patterns of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ during the first cycle at 10 mA g⁻¹ and the refined a-lattice parameters (c, charge; dc, discharge). **b**, **c**, Manganese K-edge XANES spectra from hard XAS (**b**) and O K-edge spectra from soft XAS (**c**; using total fluorescence yield mode) during the initial cycle. Features A and B are described in the text. Plots are shown for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ before cycling; 135 mAh g⁻¹ charged; 270 mAh g⁻¹ charged; 360 mAh g⁻¹ charged; 320 mAh g⁻¹ discharged after a 375 mAh g⁻¹ charge.
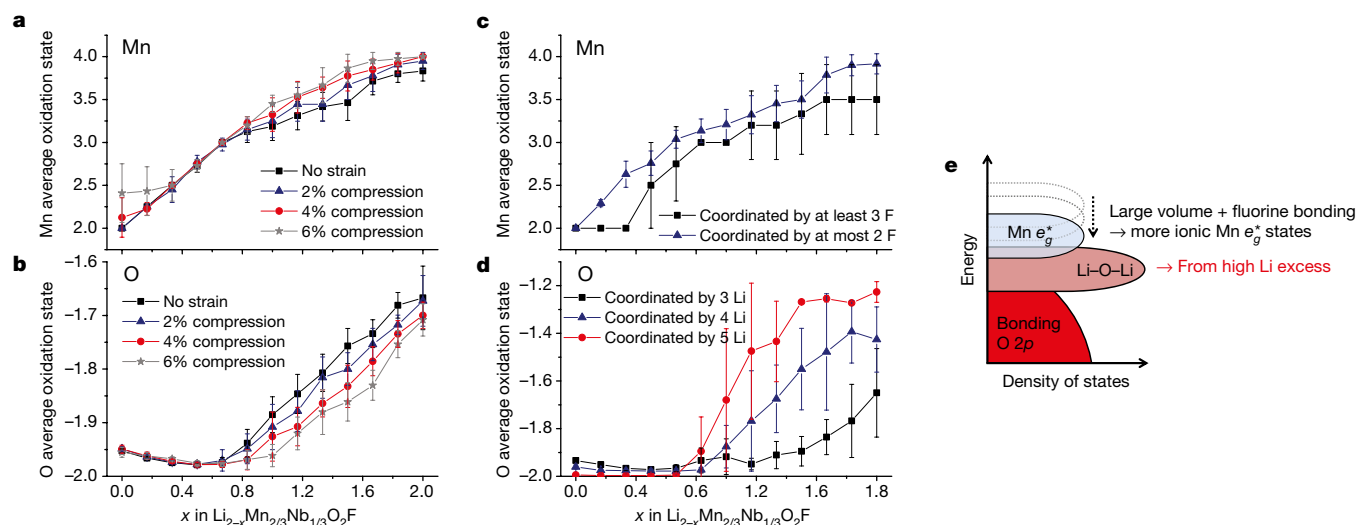
Fig. 4 | Ab initio calculations of the redox mechanism of $Li_2Mn_{2/3}Nb_{1/3}O_2F$. a, b, Manganese (**a**) and oxygen (**b**) average oxidation state as a function of delithiation (x in $Li_{2-x}Mn_{2/3}Nb_{1/3}O_2F$) and artificially introduced strain relative to the discharged state ($x = 0$). **c,** Change in the average oxidation state of Mn atoms that are coordinated by three or more fluorine atoms and those coordinated by two or fewer fluorine atoms. **d,** Change in the average oxidation state of O atoms with three, four and five Li nearest neighbours in the fully lithiated state ($x = 0$). The data in **c** and **d** were collected from model structures without strain and are representative of trends seen at all levels of strain. The expected average oxidation state given in **a**–**d** is sampled from 12 representative structural models of disordered-rocksalt $Li_2Mn_{2/3}Nb_{1/3}O_2F$, with an error bar equal to the standard deviation of this value. **e,** A schematic band structure of $Li_2Mn_{2/3}Nb_{1/3}O_2F$.

pre-edge shape and intensity, indicating Mn and O reduction. Hence, the electrochemical processes in this compound are reversible.

Ab initio study of $Li_2Mn_{2/3}Nb_{1/3}O_2F$

In conventional Li–Mn oxides (for example $LiMnO_2$, Li_2MnO_3), Mn oxidation up to Mn^{4+} is not competitive with O oxidation[8,10,20,21]. The question is then why there is a partial overlap between these redox processes in $Li_2Mn_{2/3}Nb_{1/3}O_2F$. The main differences between $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and other Mn-based Li-excess oxides are the presence of fluorine and the relatively large lattice parameter (a is about 4.26 Å for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ compared with about 4.19 Å for $Li_{1.3}Mn_{0.4}Nb_{0.3}O_2$)[8,20,21] which leads to a larger distance between Mn and the ligand. To elucidate the impact of these features on electrochemical behaviour, we study the effect of F-coordination and lattice volume on the redox mechanism using density functional theory calculations. Note that although we compute the redox mechanism through electron titration as described in the Methods section, for clarity we refer to the degree of charge in terms of Li content.

Figure 4a, b shows the Mn and O average oxidation states as a function of delithiation (x in $Li_{2-x}Mn_{2/3}Nb_{1/3}O_2F$) and for varying degrees of compressive strain relative to the fully relaxed discharged state (($l_0 - l)/l_0 \times 100\%$, where l_0 is the lattice parameter of the fully relaxed discharged state and l is the compressed lattice parameter). The vertical bars account for the range of results obtained from various structural models of disordered-rocksalt $Li_2Mn_{2/3}Nb_{1/3}O_2F$. Initial delithiation up to $x = 0.667$ (theoretical Mn^{2+}/Mn^{3+} limit) modifies only the Mn oxidation state, but further Li removal leads to concurrent O and Mn oxidation. Smaller lattice volume increases the degree of Mn^{3+}/Mn^{4+} oxidation at a fixed lithium level. At 270 mAh g^{-1} charge ($x = 1.333$), the experimentally observed lattice parameter reduction is 2.6% (Fig. 3a). At this strain, the calculations indicate average oxidation states of approximately $Mn^{3.5+}$ and $O^{1.8-}$, supporting the presence of an overlap between Mn^{3+}/Mn^{4+} and O redox, and fully consistent with our experimental results. This result seems to indicate that the large lattice parameter of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ is partly responsible for the overlap.

Figure 4c,d clarifies the impact of local environment on the oxidation of Mn and O. Figure 4c compares the average oxidation state of Mn atoms that have two or fewer F ligands out of six anion neighbours with that of Mn with three or more F ligands. At a given level

of delithiation, Mn atoms with high F-coordination are less oxidized than those with low F-coordination, indicating that the substitution of O by F favours lower Mn oxidation states and thus leads to more redox overlap with oxygen. On the oxygen side, we observe more oxidation from O atoms with five and four Li nearest neighbours than those with three Li neighbours (Fig. 4d). This trend is consistent with previous theory and experiments that indicate that the lack of transition-metal–O hybridization in Li-rich environments increases the energy of some oxygen orbitals so that they can be more easily oxidized[26–28]. Hence, the presence of Mn–F bonds, the Li-excess O environments and the larger bond distance of Mn–O(F) in this material all contribute to some competitive Mn/O oxidation at very high states of charge (Fig. 4e). Nevertheless, owing to the large Mn^{2+}/Mn^{4+} reservoir, O redox is much less needed in $Li_2Mn_{2/3}Nb_{1/3}O_2F$ than in other Mn-based Li-rich materials, rendering Mn double redox an effective way to achieve high capacity without the typical polarization and capacity fade that is observed with excessive use of the oxygen redox.

Structure and performance of $Li_2Mn_{1/2}Ti_{1/2}O_2F$

With diverse choices of high valent cations, Mn^{2+}/Mn^{4+} double redox can be realized in many different systems. As a demonstration, we have developed another new material, $Li_2Mn_{1/2}Ti_{1/2}O_2F$, in which Ti^{4+} is the high-valent cationic species. This material also forms a disordered-rocksalt phase ($a = 4.206$ Å), as can be inferred from the XRD pattern in Fig. 5a (Extended Data Table 1). SEM shows that, after mixing the compound with carbon black using a shaker-mill, the average primary particle size is about 50 nm (Fig. 5b). TEM-EDS shows a uniform distribution of Mn, Ti, O and F (Fig. 5c). As in the case of $Li_2Mn_{2/3}Nb_{1/3}O_2F$, the primary particles of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ are polycrystalline and made of grains about 15 nm across (Extended Data Fig. 9). The ^{7}Li and ^{19}F NMR results suggest the possible presence of impurities (for example LiF, Li_2O, Li_2CO_3) in our sample (Extended Data Fig. 1), but their amount is likely to be small as no crystalline or amorphous impurities could be observed with XRD and TEM (Fig. 5a, Extended Data Fig. 9).

$Li_2Mn_{1/2}Ti_{1/2}O_2F$ delivers high capacities similar to $Li_2Mn_{2/3}Nb_{1/3}O_2F$. When cycled between 1.6 V and 4.8 V (Fig. 5d) or 1.5 V and 5.0 V (Fig. 5e) at 20 mA g^{-1}, this material yields reversible capacities of 259 mAh g^{-1} (783 Wh kg^{-1}, 2,756 Wh l^{-1}) and 321 mAh g^{-1} (932 Wh kg^{-1}, 3,281 Wh l^{-1}), respectively. These values are again
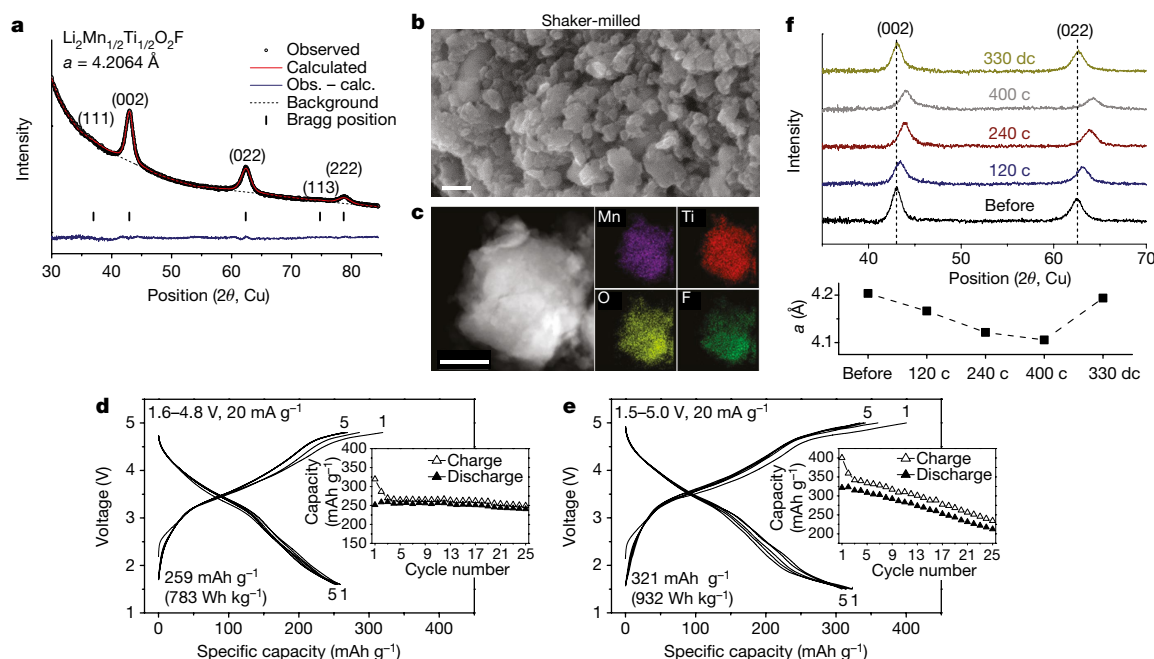
Fig. 5 | Structural characterization and electrochemical performance of $Li_2Mn_{1/2}Ti_{1/2}O_2F$. a, The XRD pattern of $Li_2Mn_{1/2}Ti_{1/2}O_2F$. **b**, SEM image of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ after shaker-milling with carbon black for electrode fabrication. Scale bar, 100 nm. **c**, EDS mapping (Mn, Ti, O, F) on a $Li_2Mn_{1/2}Ti_{1/2}O_2F$ particle. Scale bar, 200 nm. **d**, **e**, Voltage profiles and capacity retention of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ when cycled at 20 mA g^{-1} between 1.6 V and 4.8 V (**d**), and 1.5 V and 5.0 V (**e**). Cycling tests were conducted at room temperature. **f**, The XRD patterns of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ during the first cycle at 10 mA g^{-1} and the refined a-lattice parameters: before cycling; 120 mAh g^{-1} charged; 240 mAh g^{-1} charged; 400 mAh g^{-1} charged; 330 mAh g^{-1} discharged after a 400 mAh g^{-1} charge.

among the highest values achieved by advanced cathode materials[10,19]. Additional electrochemical data (rate tests, voltage window tests, change of the electrode formulation) are presented and discussed in Extended Data Fig. 10 and Methods section 'Supplementary Note 5'.

As in $Li_2Mn_{2/3}Nb_{1/3}O_2F$, the voltage profiles of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ barely change after the first cycle, indicating a reversible reaction without a major structural change or O loss, as evidenced by the DEMS results (Extended Data Fig. 5). The O-oxidation plateau appears only after charging above 230 mAh g^{-1} (above about 4.6 V), which is substantially delayed compared with other Mn-based Li-rich materials. Nevertheless, this plateau is slightly longer than in $Li_2Mn_{2/3}Nb_{1/3}O_2F$, probably owing to the smaller Mn^{2+} content in $Li_2Mn_{1/2}Ti_{1/2}O_2F$, requiring more O oxidation to achieve a given capacity.

Ex situ XRD of $Li_2Mn_{1/2}Ti_{1/2}O_2F$ indicates a reversible change in lattice parameter, shrinking from about 4.203 Å to about 4.105 Å after a 400 mAh g^{-1} charge and then recovering to about 4.194 Å after a 330 mAh g^{-1} discharge (Fig. 5f). Hard XAS confirms Mn^{2+}/Mn^{4+} redox in the material (Extended Data Fig. 11, Methods section 'Supplementary Note 6'). As in $Li_2Mn_{2/3}Nb_{1/3}O_2F$, additional capacity beyond Mn-redox capacity is probably delivered by O redox. Based on its high capacity and reversibility, $Li_2Mn_{1/2}Ti_{1/2}O_2F$ has considerable potential as a high-performance Li-ion cathode.

Outlook for Mn^{2+}/Mn^{4+} redox

Double redox couples are tremendously important for the development of advanced cathodes. Indeed, today's modern NMC-based layered cathodes all rely to some extent on the Ni^{2+}/Ni^{4+} double redox. With $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$, we have demonstrated that combined fluorination and high-valent cation substitution can introduce Mn^{2+}/Mn^{4+} redox in a Li-excess disordered-rocksalt structure, which leads to high-capacity Mn-based Li-excess cathodes (capacity of >300 mAh g^{-1}, energy density of around 1,000 Wh kg^{-1}) without an excessive use of O redox. This discovery is important, as our strategy can be widely applied to design high-performance Mn-based Li-excess cathodes that do not suffer from structural degradation triggered by extensive O redox.

The combination of Mn^{2+}/Mn^{4+} redox with the cation-disordered structure[8,13,20,22] and the partial replacement of O by F[15–17,29] leads to a large chemical space for new cathode materials. We expect to see considerable optimization through the use of different high-valent charge-compensating elements, as well as through minor alloying additions to stabilize the structure further and increase other performance aspects. The disordered-rocksalt framework has previously shown high structural stability[8,11,13,20–22], and its compositional flexibility, enabled by not requiring the preservation of the layered cathode structure, can be used to tune not only the Li-excess level for Li transport[13,14], but also the content of F and high-valent cations (such as Sn^{4+}, Sb^{5+} and Te^{6+}). These handles can all be used to modify the size of the Mn^{2+}/Mn^{4+} reservoir and balance Mn- and O-redox activities. Critical directions for further research include finding ways of decreasing the voltage slope of these compounds, so that their high capacity and energy density can be delivered over a narrower voltage window[30], as well as investigating Mn dissolution which often undermines the long-term cyclability of Mn-based cathodes[31]. Strategies based on compositional modifications of the cathode material, on changes in the short-range cation distribution, on microstructure control (for example by surface coating)[31] and on the use of tailored electrolytes[32,33] should be explored to further develop high-performance Mn^{2+}/Mn^{4+}-based cathodes for advanced Li-ion batteries.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0015-4.

1. Whittingham, M. S. Lithium batteries and cathode materials. *Chem. Rev.* **104**, 4271–4302 (2004).
2. Olivetti, E. A., Ceder, G., Gaustad, G. G. & Fu, X. Lithium-ion battery supply chain considerations: analysis of potential bottlenecks in critical metals. *Joule* **1**, 229–243 (2017).

3. Lu, Z., MacNeil, D. D. & Dahn, J. R. Layered Li[Ni$_x$Co$_{1-2x}$Mn$_x$]O$_2$ cathode materials for lithium-ion batteries. *Electrochem. Solid-State Lett* **4**, A200–A203 (2001).
4. Thackeray, M. M., Johnson, P. J., de Picciotto, L. A., Bruce, P. G. & Goodenough, J. B. Electrochemical extraction of lithium from LiMn$_2$O$_4$. *Mater. Res. Bull.* **19**, 179–187 (1984).
5. Kang, B. & Ceder, G. Battery materials for ultrafast charging and discharging. *Nature* **458**, 190–193 (2009).
6. Nitta, N., Wu, F., Lee, J. T. & Yushin, G. Li-ion battery materials: present and future. *Mater. Today* **18**, 252–264 (2015).
7. Thackeray, M. M. et al. Li$_2$MnO$_3$-stabilized LiMO$_2$ (M = Mn, Ni, Co) electrodes for lithium-ion batteries. *J. Mater. Chem.* **17**, 3112–3125 (2007).
8. Yabuuchi, N. et al. High-capacity electrode materials for rechargeable lithium batteries: Li$_3$NbO$_4$-based system with cation-disordered rocksalt structure. *Proc. Natl Acad. Sci. USA* **112**, 7650–7655 (2015).
9. Armstrong, A. R. et al. Demonstrating oxygen loss and associated structural reorganization in the lithium battery cathode Li[Ni$_{0.2}$Li$_{0.2}$Mn$_{0.6}$]O$_2$. *J. Am. Chem. Soc.* **128**, 8694–8698 (2006).
10. Hy, S. et al. Performance and design considerations for lithium excess layered oxide positive electrode materials for lithium ion batteries. *Energy Environ. Sci.* **9**, 1931–1954 (2016).
11. Freire, M. et al. A new active Li–Mn–O compound for high energy density Li-ion batteries. *Nat. Mater.* **15**, 173–177 (2016).
12. Kim, J. & Manthiram, A. A manganese oxyiodide cathode for rechargeable lithium batteries. *Nature* **390**, 265–267 (1997).
13. Lee, J. et al. Unlocking the potential of cation-disordered oxides for rechargeable lithium batteries. *Science* **343**, 519–522 (2014).
14. Urban, A., Lee, J. & Ceder, G. The configurational space of rocksalt-type oxides for high-capacity lithium battery electrodes. *Adv. Energy Mater.* **4**, 1400478 (2014).
15. Richards, W. D., Dacek, S. T., Kitchaev, D. A. & Ceder, G. Fluorination of lithium-excess transition metal oxide cathode materials. *Adv. Energy Mater.* **343**, 1701533 (2017).
16. Chen, R. et al. Disordered lithium-rich oxyfluoride as a stable host for enhanced Li$^+$ intercalation storage. *Adv. Energy Mater.* **5**, 1401814 (2015).
17. Takeda, N. et al. Reversible Li storage for nanosize cation/anion-disordered rocksalt-type oxyfluorides: LiMoO$_{2-x}$ LiF (0≤x≤2) binary system. *J. Power Sources* **367**, 122–129 (2017).
18. Urban, A., Abdellahi, A., Dacek, S., Artrith, N. & Ceder, G. Electronic-structure origin of cation disorder in transition-metal oxides. *Phys. Rev. Lett.* **119**, 176402 (2017).
19. Etcheri, V., Marom, R., Elazi, R., Salitra, G. & Aurbach, D. Challenges in the development of advanced Li-ion batteries: a review. *Energy Environ. Sci.* **4**, 3243–3262 (2011).
20. Wang, R. et al. A disordered rock-salt Li-excess cathode materials with high capacity and substantial oxygen redox activity: Li$_{1.25}$Nb$_{0.25}$Mn$_{0.5}$O$_2$. *Electrochem. Commun.* **60**, 70–73 (2015).
21. Yabuuchi, N. et al. Origin of stabilization and destabilization in solid-state redox reaction of oxide ions for lithium-ion batteries. *Nat. Commun.* **7**, 13814 (2016).
22. Lee, J. et al. A new class of high capacity cation-disordered oxides for rechargeable lithium batteries: Li–Ni–Ti–Mo oxides. *Energy Environ. Sci.* **8**, 3255–3265 (2015).
23. Manceau, A., Marcus, M. A. & Grangeon, S. Determination of Mn valence states in mixed-valent manganates by XANES spectroscopy. *Am. Mineral.* **97**, 816–827 (2012).
24. de Groot, F. M. F. et al. Oxygen 1s X-ray-absorption edges of transition-metal oxides. *Phys. Rev. B* **40**, 5715–5723 (1989).
25. Kurata, H., Lefèvre, E. & Colliex, C. Electron-energy-loss near-edge structures in the oxygen K-edge spectra of transition-metal oxides. *Phys. Rev. B* **47**, 13763–13768 (1993).
26. Luo, K. et al. Charge-compensation in 3d-transition-metal-oxide intercalation cathodes through the generation of localized electron holes on oxygen. *Nat. Chem.* **8**, 684–691 (2016).
27. Seo, D.-H. et al. The structural and chemical origin of the oxygen redox activity in layered and cation-disordered Li-excess cathode materials. *Nat. Chem.* **8**, 692–697 (2016).
28. Xie, Y., Saubanère, M. & Doublet, M.-L. Requirements for reversible extra-capacity in Li-rich layered oxides for Li-ion batteries. *Energy Environ. Sci.* **10**, 266–274 (2017).
29. Lee, J. et al. Mitigating oxygen loss to improve the cycling performance of high capacity cation-disordered cathode materials. *Nat. Commun.* **8**, 981 (2017).
30. Abdellahi, A., Urban, A., Dacek, S. & Ceder, G. Understanding the effect of cation disorder on the voltage profile of lithium transition-metal oxides. *Chem. Mater.* **28**, 5373–5383 (2016).
31. Cho, J., Kim, T.-J., Kim, Y. J. & Park, B. Complete blocking of Mn^{3+} ion dissolution from a LiMn$_2$O$_4$ spinel intercalation compound by Co$_3$O$_4$ coating. *Chem. Commun.* 1074–1075 (2001).
32. Xu, K. Electrolytes and interphases in Li-ion batteries and beyond. *Chem. Rev.* **114**, 11503–11618 (2014).
33. Wang, Y. et al. Design principles for solid-state lithium superionic conductors. *Nat. Mater.* **14**, 1026–1031 (2015).

METHODS

Synthesis. To synthesize $Li_2Mn_{2/3}Nb_{1/3}O_2F$, we used Li_2O (Sigma-Aldrich, 97%), MnO (Alfa Aesar, 99%), Nb_2O_5 (Alfa Aesar, 99.9%) and LiF (Alfa Aesar, 99.99%) as precursors. For $Li_2Mn_{1/2}Ti_{1/2}O_2F$, we used Li_2O (Sigma-Aldrich, 97%), MnO (Alfa Aesar, 99%), TiO_2 (Alfa Aesar, 99.9%) and LiF (Alfa Aesar, 99.99%) as precursors. Other than Li_2O, for which we used 10% excess (rather than stoichiometric amount) to compensate for possible loss of Li_2O during synthesis, stoichiometric amounts of precursors were dispersed into (Ar-filled) stainless steel jars and then planetary ball-milled (Retch PM 200) for 40 h at the rate of 450 r.p.m., during which $Li_2Mn_{2/3}Nb_{1/3}O_2F$ or $Li_2Mn_{1/2}Ti_{1/2}O_2F$ form mechanochemically. The total amount of precursors in each jar (50 ml) was approximately 1 g, and five 10-mm-diameter and ten 5-mm-diameter stainless balls were used as the grinding media.

Electrochemistry. To prepare a cathode film from $Li_2Mn_{2/3}Nb_{1/3}O_2F$ or $Li_2Mn_{1/2}Ti_{1/2}O_2F$, 480 mg of active compounds and 240 mg of carbon black (Timcal, SUPER C65) were first mixed for an hour in an Ar-filled 45-ml zirconia vial with 10 g of 5-mm-diameter yttria-stabilized zirconia balls (Inframat Advanced Materials) as grinding media, using a SPEX 8000M Mixer/Mill. Polytetrafluoroethylene (PTFE, DuPont, Teflon 8 A) was then added to the mixture as a binder, such that the cathode film consists of the active compounds, carbon black and PTFE in the weight ratio of 60:30:10. The weight ratio for cathode films with higher active-material loading was either 70:20:10 or 80:15:5. The components were then manually mixed using a mortar and pestle and rolled into a thin film inside an Ar-filled glove box. To assemble a cell for all cycling tests, 1 M of $LiPF_6$ in ethylene carbonate and dimethyl carbonate (EC/DMC) solution (1:1, BASF), glass microfibre filters (Whatman) and Li metal foil (FMC) were used as the electrolyte, the separator and the counter electrode, respectively. Coin cells (CR2032) were assembled in an Ar-filled glove box and tested on a Maccor 2200 or an Arbin battery cycler at room temperature in the galvanostatic mode otherwise specified. The loading density of the cathode film was about $6\,mg\,cm^{-2}$. The specific capacity was calculated on the amount of the active compounds in the cathode film.

Characterization. XRD patterns of the as-prepared compounds and electrodes were collected on a Rigaku MiniFlex diffractometer (Cu source) in the 2θ range of 5°–85°. Rietveld refinement was completed with PANalytical X'pert HighScore Plus software. Elemental analysis on the compounds by Luvak Inc. was performed with direct current plasma emission spectroscopy (ASTM E 1097-12) for Li, Mn, Nb and Ti, and with an ion-selective electrode (ASTM D1179-10) for F. SEM images were collected on a Zeiss Gemini Ultra-55 Analytical Field Emission SEM in the Molecular Foundry at Lawrence Berkeley National Laboratory (LBNL). For TEM sampling, particles were sonicated with ethanol and drop-cast on an ultrathin carbon grid. Scanning TEM/EDS spectra were acquired from a few of the particles on a JEM-2010F microscope equipped with an X-max EDS detector in the Molecular Foundry at LBNL.

Hard X-ray absorption spectroscopy. We performed Mn, Nb and Ti K-edge XANES measurements in transmission mode using beamline 20BM at the Advanced Photon Source. The incident energy was selected using a Si(111) monochromator. We performed the energy calibration by simultaneously measuring the spectra of the appropriate metal foil. Harmonic rejection was accomplished using a Rh-coated mirror. The samples for the measurements were prepared with the $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ electrode films before and after first charging and discharging to designated capacities. The loading density of the films was approximately $10\,mg\,cm^{-2}$. Additionally, we measured the spectra of some reference standards in transmission mode, to aid interpretation of the XANES data. Data reduction was carried out using the Athena software[34].

Soft X-ray absorption spectroscopy. Soft XAS measurements on the O K-edge were performed on Beamline 8.0.1.1 (iRIXS endstation) at the Advanced Light Source, LBNL[35]. All the O K-edge XAS spectra were normalized by incident beam flux monitored by a gold mesh, which was located in front of the ultra-high-vacuum experimental chamber. The energy resolution of the O K-edge XAS spectra was set to 0.2 eV and a reference (anatase) TiO_2 O K-edge XAS spectrum was also recorded for careful energy calibration during the XAS experiments. XAS spectra taken in total fluorescence yield mode were chosen and presented as Fig. 3c in this paper to represent the bulk-like information (typically a few hundred nanometres below the sample surface) from these cathode materials. In addition, sample preparation and handling for X-ray spectroscopy were done in an air-free environment to avoid surface contamination and oxidation.

Differential electrochemical mass spectrometer measurement. A DEMS measurement was used to detect and quantify O_2 and CO_2 gas evolved during charging and discharging (Extended Data Fig. 5). The custom-built DEMS and the cell geometry used are described in previous publications[36–38]. The electrochemical cells used with the DEMS device were prepared in a dry Ar glove box (<1 ppm O_2 and H_2O, MBraun USA, Inc.) using the modified Swagelok design and the same materials as discussed previously. The assembled cells were charged under a static head of positive Ar pressure (approximately 1.2 bar) after being appropriately attached to the DEMS. Throughout the charge, Ar gas pulses periodically swept accumulated gases to a mass spectrometer chamber. The mass spectrometer absolute sensitivity has been calibrated for CO_2 and O_2, and therefore the partial pressures of these gases can be determined. The amount of CO_2 and O_2 evolved is then quantified based on the volume of gas swept to the mass spectrometer per pulse.

Solid-state NMR spectroscopy. We acquired all ^7Li and ^{19}F NMR data at room temperature on a Bruker Avance500 WB spectrometer (11.7 T), at Larmor frequencies of -194.4 and -70.7 MHz, respectively. The data were obtained under 50-kHz magic-angle spinning (MAS), using a 1.3-mm double-resonance probe. The chemical shifts of ^7Li and ^{19}F were referenced against lithium fluoride powder (LiF, $\delta_{iso}(^7Li) = -1$ ppm and $\delta_{iso}(^{19}F) = -204$ ppm). ^7Li spin echo spectra were acquired on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ using a 90° radio-frequency (RF) excitation pulse of 0.9 μs and a 180° RF pulse of 1.8 μs at 110 W. A recycle delay of 0.03 s was found to be sufficiently long to ensure complete relaxation of all Li signals between the excitation pulses. Lineshape analysis was carried out using the SOLA lineshape simulation package within the Bruker TOPSPIN software. Because the resonant frequency range of the ^{19}F nuclei in the as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ samples is larger than the excitation bandwidth of the RF pulse used in the NMR experiment, nine spin echo spectra were collected on each sample, with the irradiation frequency varied in steps equal to the excitation bandwidth of the RF pulse (330 ppm or 155 kHz). The individual sub-spectra were processed using zero-order phase correction so that the on-resonance signal was in the absorption mode. The nine sub-spectra were then added to give an overall sum spectrum with no further phase correction required. We note that this methodology, termed 'spin echo mapping'[39], 'frequency stepping'[40,41] or 'VOCS' (variable offset cumulative spectrum)[42], is required to provide a large excitation bandwidth and uniformly excite the broad F signals. Individual ^{19}F spin echo spectra were collected using a 90° RF excitation pulse of 1.6 μs and a 180° RF pulse of 3.2 μs at 76.3 W (or 156 kHz), with a recycle delay of 0.05 s. For comparison, a spin echo spectrum was collected on LiF using similar RF pulses but a longer recycle delay of 14 s. A ^{19}F spin echo spectrum, acquired under the same conditions as the $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ spin echo spectra but on an empty rotor, revealed no significant background signal coming from the NMR probe itself.

Density functional theory calculations. Density functional theory analysis of the redox mechanism was performed with the Vienna Ab-Initio Simulation Package (VASP)[43] using the projector augmented-wave method. First, structural models of the $Li_2Mn_{2/3}Nb_{1/3}O_2F$ disordered rocksalt were obtained using a cluster-expansion-based Monte Carlo simulation, chosen to find low-energy disordered structures with representative short-range order, while suppressing phase separation. The cluster-expansion Hamiltonian used for the Monte Carlo simulations consists of a decomposition of the internal energy of a particular atomic configuration on a rocksalt lattice into contributions from two-, three-, and four-body terms up to maximum interaction distance of 7.0 Å, 4.1 Å and 4.1 Å respectively, relative to an ideal rocksalt lattice with a primitive lattice constant of 3.0 Å, on top of an electrostatic model based on the formal charges of all species[44]. To obtain the interaction terms, we first calculated 450 representative configurations of Li^+, Mn^{2+}, Nb^{5+}, O^{2-} and F^- on a rocksalt lattice within the Perdew–Burke–Ernzerhof exchange-correlation functional[45], supplemented with the rotationally invariant Hubbard U correction[46] to the transition metal d states to correct self-interaction error ($U_{Mn} = 3.9$ eV, $U_{Nb} = 1.5$ eV based on previously reported fits to oxide formation enthalpies[47]). These calculations were performed with a reciprocal-space discretization of $25\,Å^{-1}$, 520 eV plane-wave cut-off, and a 10^{-5} eV and $0.02\,eV\,Å^{-1}$ convergence on total energy and interatomic forces respectively. The strength of each cluster interaction, as well as the dielectric constant, was then fitted using a L_1-regularized least-squares regression, optimized by cross-validation, which resulted in an out-of-sample error of 9 meV per atom.

The redox mechanism of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ was calculated on 12 structural models of 36 atoms each, obtained from the Monte Carlo simulations described above. Oxidation calculations were done using the hybrid Heyd–Scuseria–Ernzerhof functional[48], using a 650-eV plane-wave cut-off, $10\,Å^{-1}$ reciprocal-space discretization, and a 10^{-5} eV and $0.02\,eV\,Å^{-1}$ convergence on total energy and interatomic forces respectively. The fraction of exact exchange was set to 0.30 on the basis of a calibration to the Kohn–Sham gaps of α-Mn_3O_4, γ-MnOOH and β-MnO_2 calculated within the G_0W_0 approximation, following previously reported methodology for reproducing the redox competition between transition metals and oxygen[49]. To investigate the order in which various redox couples are activated in the material, suppressing major structural rearrangements, we trace the oxidation state of each species (obtained from the magnetic moment of each atom) as electrons are removed from the material and charge compensated by a uniform background charge[50], allowing the local atomic arrangements to relax at each step but keeping the lattice fixed. As the order of oxidation reactions is determined by the character of the valence-band edge at various states of charge, such electron titration provides an efficient way to look at the electronic contribution to the redox mechanism.

Supplementary Note 1. The absence of peaks other than those corresponding to the disordered-rocksalt phases in the XRD data collected on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ suggests that the samples are fairly phase-pure without large amounts of crystalline impurities such as LiF, Li_2O, or Li_2CO_3. Nevertheless, because small amounts of impurity phases or amorphous phases can be invisible to XRD, we investigated further using ^7Li and ^{19}F NMR spectroscopy.

^7Li spin echo NMR spectra, obtained on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ samples, are shown in Extended Data Fig. 1a, b. The data have been fitted using a minimum of three Li sites: Li1, Li2 and Li3. The fits suggest about 78%–79% ($\pm$2%) of Li in paramagnetic environments (Li1 and Li2 signals), and about 21%–22% ($\pm$2%) Li in diamagnetic environments (Li3 signal). The broad Li1 and Li2 resonances are ascribed to several paramagnetic Li environments close to open-shell Mn and with very similar shifts, resulting in overlapping signals. Paramagnetic interactions between unpaired Mn d electrons and the Li nuclei lead to a broadening of the individual Li signals with shifts centred around 64.9 ppm (Li1) and -27.4 ppm (Li2) for $Li_2Mn_{1/2}Ti_{1/2}O_2F$, and around 57.7 ppm (Li1) and -25.5 ppm (Li2) for $Li_2Mn_{2/3}Nb_{1/3}O_2F$. The sharper Li3 resonance, with a shift close to 0 ppm (-0.1 ppm for $Li_2Mn_{1/2}Ti_{1/2}O_2F$ and -0.5 ppm for $Li_2Mn_{2/3}Nb_{1/3}O_2F$), is ascribed to diamagnetic Li sites in the samples. Because Li^+, Nb^{5+} and Ti^{4+} do not have unpaired electrons, Li nuclei in diamagnetic Li/Ti- and Li/Nb-rich domains in $Li_2Mn_{1/2}Ti_{1/2}O_2F$ and $Li_2Mn_{2/3}Nb_{1/3}O_2F$ have a shift around 0 ppm that cannot be distinguished from that of potential Li_2O, LiF and Li_2CO_3 impurities, resonating at 2.8, -1 and 0 ppm, respectively[51]. All of these Li environments may contribute to the Li3 signal, and individual contributions cannot be quantified. In fact, local segregation of cations that would lead to Li/Ti- or Li/Nb-rich domains in our compounds has been observed in several compounds—for example, Li_2MnO_3-like domains in Li- and Mn-rich layered Ni–Mn–Co materials[10], or Li_3NbO_4-like local domains in disordered Li–V–Nb–O materials[52]. A previous ^7Li NMR study on paramagnetic Li transition metal phosphates ($LiMPO_4$) cathodes has found that paramagnetic shift contributions from distant M beyond the second metal coordination shell around the central Li can be non-zero[53]. This observation suggests that, in $Li_2Mn_{1/2}Ti_{1/2}O_2F$ and $Li_2Mn_{2/3}Nb_{1/3}O_2F$, Mn is likely to be more than 7 Å away from the Li for there to be no paramagnetic shift contribution and an overall Li shift close to 0 ppm.

^{19}F spin echo sum spectra, collected on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$, are compared to the spin echo spectrum collected on crystalline LiF powder and presented in the Extended Data Fig. 1c. Further details on how the sum spectra were obtained can be found in the Methods section for solid state NMR spectroscopy where we describe the method of 'spin echo mapping'. The ^{19}F NMR data clearly indicate that most of the F is found in paramagnetic environments (that is, with Mn in the first, second and/or third metal coordination shell around the F nucleus), giving rise to very broad overlapping NMR signals shifted away from the LiF resonant frequency. Nevertheless, LiF-like F environments are also observed as a sharp signal with a resonant frequency equal to that of pure LiF (-204 ppm). Some of our current work on similar paramagnetic cation-disordered oxyfluorides suggests that F nuclei directly bonded to the paramagnetic centre (here Mn) are essentially invisible in the NMR spectrum, because the very strong interaction with the unpaired electrons leads to extremely broad resonances with a very large shift that are lost in the background. Hence, we suspect our ^{19}F NMR data not to be quantitative and the proportion of F in paramagnetic environments to be even larger than that determined from experimental observations. With this in mind, the ^{19}F NMR data confirm that most of the F has integrated into the bulk cation-disordered oxide lattice. Although the -204 ppm ^{19}F signal can indicate LiF impurity in our samples, it can also indicate the presence of a small proportion of LiF-like domains in the disordered oxyfluoride structure, which would be consistent with recent theory work[15] showing that the much higher energetic cost of creating M–F bonds, as compared with Li–F bonds, results in the incorporation of F in Li-rich (that is, LiF-like) local environments in cation-disordered oxyfluoride materials.

In short, diamagnetic Li sites and LiF-like F environments observed with NMR cannot be uniquely attributed to either local domains in the rocksalt structure or to amorphous impurity phases, such as LiF, Li_2O or Li_2CO_3, in our samples. Hence, NMR can give us an upper bound to the amount of impurity present in the samples but does not enable us to obtain the exact amount of potential impurity phases. In the extreme case in which all of the diamagnetic Li and LiF-like F signals come from Li_2O and LiF, the total weight fraction of impurity phases is estimated to be no more than 6–7 wt%; it is likely to be less, as no crystalline impurities were observed with XRD and no amorphous domains were observed in TEM (Extended Data Figs. 2, 9). As a result, we are confident stating that the performance of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ is predominantly determined by the transition metal oxyfluoride rocksalt phase.

Supplementary Note 2. Extended Data Fig. 5a, b shows the O_2 (g) and CO_2 (g) evolution data from $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ during initial charge (1.5–5.0 V, 20 mA g^{-1}), collected by DEMS measurements. The capacity observed

during this DEMS test is slightly smaller than that in a coin cell test, because the electrode films were made thicker (about 13 mg cm^{-2} versus about 6 mg cm^{-2} in coin cells) for this measurement to enhance gas evolution signals. We detect negligible O_2 (g) evolution from both compounds upon first charging to 5 V. The total amount of O_2 (g) evolved during the first charge is smaller than 0.01 μmol mg^{-1} (of active material) for both $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$, which corresponds to less than 0.2% of total oxygen content in the two materials. For conventional layered Li- and Mn-rich cathodes, such as $Li_{1.2}Ni_{0.13}Co_{0.13}Mn_{0.54}O_2$, oxygen loss occurs dominantly in the form of O_2 (g) evolution which starts from above 4.5 V in the first charge, and results in a loss of about 4%–5% of the total oxygen content of the cathode materials[10,26]. Therefore, the remarkably small amount of O_2 (g) evolved even until 5 V demonstrates negligible oxygen loss from both $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ compounds.

Interestingly, we detect a noticeable amount of CO_2 (g) evolved from the two materials (0.30 μmol mg^{-1} and 0.24 μmol mg^{-1} for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$, respectively), with much of the evolved CO_2 coming at lower voltages than the threshold voltage (about 4.5 V) for decomposition of 1 M $LiPF_6$ in EC/DMC electrolyte[32]. On the basis of an acid titration test using 1 M H_2SO_4[54,55], we find that most of this CO_2 (g) is likely to come from electrochemical decomposition of surface carbonates (for example solid lithium carbonate) that probably form during the shaker-milling process between the active compounds and carbon black. For instance, Extended Data Fig. 5c shows the cumulative CO_2 evolution during acid titration on shaker-milled $Li_2Mn_{1/2}Ti_{1/2}O_2F$ and carbon black mixture. CO_2 (g) evolves from the mixture immediately after adding 1 M H_2SO_4, with a total CO_2 amount of about 0.17 μmol per mg of $Li_2Mn_{1/2}Ti_{1/2}O_2F$. This direct CO_2 evolution indicates a chemical decomposition of an equimolar amount of carbonate species by the H_2SO_4 addition (about 0.7 wt% of the powder mixture, assuming Li_2CO_3 as the carbonate species), which can also decompose electrochemically. Since 5 V is a high enough voltage to electrochemically decompose carbonates[32,55], we expect that a similar amount of CO_2 (g) to that in the acid titration would evolve from the surface carbonates during charging of the $Li_2Mn_{1/2}Ti_{1/2}O_2F$ electrode, which would imply that about 71% (about 0.17 μmol mg^{-1} out of 0.24 μmol mg^{-1}) of the CO_2 evolved during the first charge originates from carbonate decomposition.

It is worth noting that some sub-surface Li_2CO_3 may not be detected using our acid titration method, although this carbonate may still oxidize to CO_2 during the first charge cycle such that carbonate oxidation accounts for nearly all CO_2 evolution observed in Extended Data Fig. 5a, b. From the many transition metal oxides studied using our gas evolution methods (for example, Ni-rich and Li/Mn-rich Ni–Mn–Co oxides)[55], residual Li_2CO_3, and not electrolyte degradation (below 4.8 V), has accounted for all CO_2 evolution during the first charge cycle, and it is likely that a similar phenomenon is observed for these materials. Nevertheless, it is possible, although less likely, that the additional CO_2 evolved beyond that expected from the titrated Li_2CO_3 quantity may come from direct electrolyte decomposition, particularly at high voltages (>4.8 V), or from some oxygen species evolved from the materials reacting with the electrolyte[26,32,55,56].

Supplementary Note 3. Most of the Li-excess Mn-rich cathodes using high levels of oxygen redox experience voltage fading, a continuous reduction of both charge and discharge voltages upon extended cycling[10,57]. From the evolution of average voltages upon cycling (Extended Data Fig. 6), we find that voltage fading for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ is less pronounced than for other Mn-rich cathodes. Comparing the 2nd and 20th cycles between 1.5 and 4.6 V, 1.5 and 4.8 V, and 1.5 and 5.0 V at 20 mA g^{-1}, we observe a decrease of the average discharge voltage by approximately 1.3%, 2.2% and 4.0%, respectively. Apparently, a higher charge cut-off voltage results in more reduction of discharge voltage upon cycling. On the contrary, the average charge voltage increases by about 1.8%, 1.3%, and 1.9%, when comparing the 2nd and 20th cycles between 1.5 and 4.6 V, 1.5 and 4.8 V, and 1.5 and 5.0 V, respectively. In fact, half of the average charge–discharge voltages ((charge voltage + discharge voltage)/2) changes only about 0.3%, 0.27% and 0.7%, when comparing the 2nd and 20th cycles between 1.5 and 4.6 V, 1.5 and 4.8 V, and 1.5 and 5.0 V, respectively. For the Mn-rich Li-excess materials that experience voltage fading, both the discharge and charge voltages decrease upon cycling[10,57]. Our result, on the other hand, shows slight decrease of discharge voltage but increase of charge voltage, and the average of the two barely changes. This indicates that the voltage change for $Li_2Mn_{2/3}Nb_{1/3}O_2F$ is unlikely to be the result of irreversible voltage fading but is the result of impedance growth such as from electrolyte decomposition at high voltages above 4.5 V[32].

Supplementary Note 4. We performed hard XAS on the $Li_2Mn_{2/3}Nb_{1/3}O_2F$ material. Along with rising edges (Extended Data Fig. 7a), pre-edge features of XANES spectra (Extended Data Fig. 7b) can give information about oxidation states. The pre-edge feature (about 6,539 eV) in the Mn K-edge XANES spectra originates from electron excitation from the Mn $1s$ state to mixed Mn $3d$–$4p$ states, allowed in a non-centrosymmetric environment[58]. Direct comparison of the Mn K-edge pre-edge features of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ upon cycling is shown in Extended Data Fig. 7b. To analyse their shape more clearly, first derivatives of their pre-edges are

shown in Extended Data Fig. 7c–e. The first derivatives of the spectra from 'before cycle' and 'after first charging to 375 mAh g^{-1} then discharging to 320 mAh g^{-1}' resemble that of MnO (Extended Data Fig. 7c), suggesting that most Mn ions in the two samples are in the Mn^{2+} state. After first charging Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F to 135 mAh g^{-1}, the derivative shape looks similar to that of Mn$_2$O$_3$, indicating the presence of Mn^{3+} in the sample (Extended Data Fig. 7d). After charging to 270 and 360 mAh g^{-1}, the derivative shape changes towards that of MnO$_2$, indicating mostly Mn^{4+} ions although Mn^{3+} and Mn^{2+} might also be present (Extended Data Fig. 7e).

Supplementary Note 5. Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F exhibits promising cycling behaviour, as does Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F. When cycled between 1.6 V and 5.0 V (Extended Data Fig. 10a), 2.0 V and 4.8 V (Extended Data Fig. 10b), and 2.3 V and 4.6 V (Extended Data Fig. 10c) at 20 mA g^{-1}, the 60 wt%:30 wt%:10 wt% = Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F: carbon black:PTFE electrode delivers discharge capacities up to 306 mAh g^{-1} (920 Wh kg^{-1}), 227 mAh g^{-1} (739 Wh kg^{-1}), and 160 mAh g^{-1} (534 Wh kg^{-1}), respectively. Rate capability of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F is acceptable. When cycled at high rates of 200 and 400 mA g^{-1} between 1.6 V and 5.0 V, the material still delivers discharge capacities up to 210 mAh g^{-1} (629 Wh kg^{-1}) and 158 mAh g^{-1} (461 Wh kg^{-1}) (Extended Data Fig. 10d). Capacity retention of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F (Extended Data Fig. 7e) is good and is slightly better than that of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F (Extended Data Fig. 4). When cycled at 100 mA g^{-1} and above, the capacity loss during initial 25 cycles is less than 0.4% per cycle. The 80 wt%:15 wt%:5 wt% = Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F:carbon black:PTFE electrode exhibits similar performance to the 60 wt%:30 wt%:10 wt% = Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F:carbon black:PTFE electrode (Extended Data Fig. 10b, f).

Supplementary Note 6. Extended Data Fig. 11a, b shows the Mn K-edge XANES spectra of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F before cycle, after charging to 120 mAh g^{-1}, 240 mAh g^{-1} and 400 mAh g^{-1}, and after charging to 400 mAh g^{-1} then discharging to 330 mAh g^{-1}. Upon first charging from 0 to 120 mAh g^{-1} and 240 mAh g^{-1}, the Mn rising-edge shifts, from an energy in between those in MnO and Mn$_2$O$_3$, to an energy in Mn$_2$O$_3$ and then partway up to an energy in MnO$_2$. Further change is small upon charging to 400 mAh g^{-1}. The edge returns to the original position after discharging to 330 mAh g^{-1}. This result suggests that Mn ions in the as-prepared Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F compounds are mostly Mn^{2+} (possibly with some Mn^{3+}), which are oxidized in charge towards Mn^{4+} with some Mn ions not fully oxidized. Upon discharge, Mn ions return to Mn^{2+}. Note that because the shape of Mn K-edge spectra for a given oxidation state can vary a lot depending on bonding environment[58], and there are no reported references for Mn-based disordered-oxyfluoride compounds, quantitative analysis of our results is difficult.

Derivative analysis on the Mn pre-edge feature at about 6,539 eV (Extended Data Fig. 11c–e) suggests the same Mn-redox mechanism. The first derivatives of the spectra from 'before cycle' and 'after first charging to 400 mAh g^{-1} then discharging to 330 mAh g^{-1}' exhibit a mixed shape of the first-derivative spectra of MnO and Mn$_2$O$_3$ (Extended Data Fig. 11c). This suggests an existence of Mn^{2+} ions with some partly oxidized Mn ions such as Mn^{3+}. After first charging to 120 mAh g^{-1}, the derivative shape looks similar to that of Mn$_2$O$_3$, indicating Mn^{3+} in the sample (Extended Data Fig. 11d). After charging to 240 and 400 mAh g^{-1}, the derivative shape changes towards that of MnO$_2$, suggesting a large amount of Mn^{4+} ions, but Mn^{3+} and Mn^{2+} might also be present (Extended Data Fig. 11e).

The Ti K-edge spectra of the Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F samples (Extended Data Fig. 11f) resemble that of TiO$_2$ (Ti^{4+}) and their rising-edge position barely changes during cycling, although there are minor changes in shape, which indicates local Ti-site distortion[59]. This suggests that Ti exists as Ti^{4+} and is redox-silent. Because Ti^{4+} is redox inactive, we expect that reversible capacities of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F beyond Mn capacities come from O redox, as in the case of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F.

Data availability. The datasets generated and analysed during this study are available from the corresponding authors on reasonable request.

34. Ravel, B. & Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *J. Synchrotron Radiat.* **12**, 537–541 (2005).

35. Yang, W. et al. Key electronic states in lithium battery materials probed by soft X-ray spectroscopy. *J. Electron Spectrosc. Relat. Phenom.* **190**, 64–74 (2013).
36. McCloskey, B. D., Bethune, D. S., Shelby, R. M., Girishkumar, G. & Luntz, A. C. Solvents' critical role in nonaqueous lithium–oxygen battery electrochemistry. *J. Phys. Chem. Lett.* **2**, 1161–1166 (2011).
37. McCloskey, B. D. et al. On the efficacy of electrocatalysis in nonaqueous Li-O$_2$ batteries. *J. Am. Chem. Soc.* **133**, 18038–18041 (2011).
38. McCloskey, B. D. et al. Twin problems of interfacial carbonate formation in nonaqueous Li-O$_2$ batteries. *J. Phys. Chem. Lett.* **3**, 997–1001 (2012).
39. Sananes, M. T., Tuel, A., Hutchings, G. J. & Volta, J. C. Characterization of different precursors and activated vanadium phosphate catalysis by ^{31}P NMR spin echo mapping. *J. Catal.* **148**, 395–398 (1994).
40. O'Dell, L. A. & Schurko, R. W. QCPMG using adiabatic pulses for faster acquisition of ultra-wideline NMR spectra. *Chem. Phys. Lett* **464**, 97–102 (2008).
41. Pell, A. J., Clément, R. J., Grey, C. P., Emsley, L. & Pintacuda, G. Frequency-stepped acquisition in nuclear magnetic resonance spectroscopy under magic angle spinning. *J. Chem. Phys.* **138**, 114201 (2013).
42. Massiot, D. et al. ^{71}Ga and ^{69}Ga nuclear magnetic resonance study of β-Ga$_2$O$_3$: resolution of four- and six-fold coordinated Ga sites in static conditions. *Solid State Nucl. Magn. Reson.* **4**, 241–248 (1995).
43. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
44. Richards, W. D., Wang, Y., Miara, L., Kim, J. C. & Ceder, G. Design of Li$_{1+2x}$Zn$_{1-x}$PS$_4$, a new lithium ion conductor. *Energy Environ. Sci.* **9**, 3272–3278 (2016).
45. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
46. Dudarev, S., Botton, G. A., Savrasov, S. Y. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. *Phys. Rev. B* **57**, 1505–1509 (1998).
47. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA+U framework. *Phys. Rev. B* **73**, 8627 (1994).
48. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *J. Chem. Phys.* **118**, 8207 (2003).
49. Seo, D.-H., Urban, A. & Ceder, G. Calibrating transition-metal energy levels and oxygen bands in first-principles calculations: accurate prediction of redox potentials and charge transfer in lithium transition-metal oxides. *Phys. Rev. B* **92**, 115118 (2015).
50. Raebiger, L. S. & Zunger, A. Charge self-regulation upon changing the oxidation state of transition metals in insulators. *Nature* **453**, 763–766 (2008).
51. Dupré, N., Cuisinier, M. & Guyomard, D. Electrode/electrolyte interface studies in lithium batteries using NMR. *Electrochem. Soc. Interface* **20**, 61–67 (2011).
52. Nakajima, M. & Yabuuchi, N. Lithium-excess cation-disordered rocksalt-type oxide with nanoscale phase segregation: Li$_{1.25}$Nb$_{0.25}$V$_{0.5}$O$_2$. *Chem. Mater.* **29**, 6927–6935 (2017).
53. Middlemiss, D. S., Iliott, A. J., Clément, R. J., Strobridge, F. C. & Grey, C. P. Density functional theory-based bond pathway decompositions of hyperfine shifts: equipping solid-state NMR to characterize atomic environments in paramagnetic materials. *Chem. Mater.* **25**, 1723–1734 (2013).
54. McCloskey, B. D. et al. Combining accurate O$_2$ and Li$_2$O$_2$ assays to separate discharge and charge stability limitations in nonaqueous Li–O$_2$ batteries. *J. Phys. Chem. Lett.* **4**, 2989–2993 (2013).
55. Renfrew, S. & McCloskey, B. D. Residual lithium carbonate predominantly accounts for first cycle CO$_2$ and CO outgassing of Li-stoichiometric and Li-rich layered transition-metal oxides. *J. Am. Chem. Soc.* **139**, 17853–17860 (2017).
56. Yabuuchi, N., Yoshii, J., Myung, S.-T., Nakai, I. & Komaba, S. Detailed studies of a high-capacity electrode material for rechargeable batteries, Li$_2$MnO$_3$– LiCo$_{1/3}$Ni$_{1/3}$Mn$_{1/3}$O$_2$. *J. Am. Chem. Soc.* **133**, 4404–4419 (2011).
57. Bettge, M. et al. Voltage fade of layered oxides: its measurement and impact on energy density. *J. Electrochem. Soc.* **160**, A2046–A2055 (2013).
58. Chalmin, E., Farges, F. & Brown, G. E. Jr. A pre-edge analysis of Mn K-edge XANES spectra to help determine the speciation of manganese in minerals and glasses. *Contrib. Mineral. Petrol.* **157**, 111–126 (2009).
59. Farges, F., Brown, G. E. Jr & Rehr, J. J. Coordination chemistry of Ti(IV) silicate glasses and melts: XAFS study of titanium coordination in oxide model compounds. *Geochim. Cosmochim. Acta* **60**, 3023–3038 (1996).
60. Piilonen, P. C. et al. Structural environment of Nb^{5+} in dry and fluid-rich (H$_2$O, F) silicate glasses: a combined XANES and EXAFS study. *Can. Mineral.* **44**, 775–794 (2006).
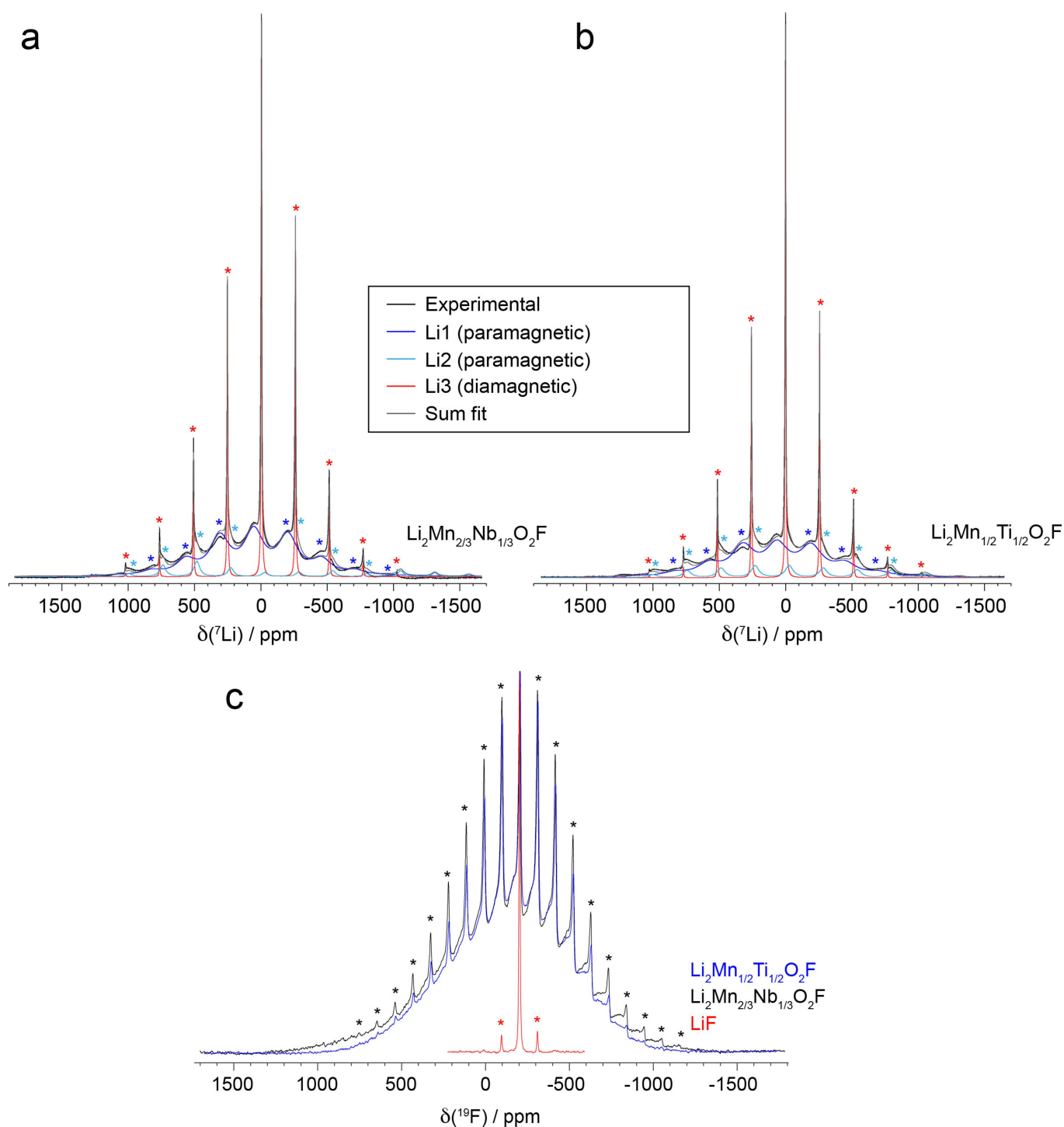
**Extended Data Fig. 1 | Solid-state NMR spectroscopy results. a, b,** ^{7}Li spin echo NMR spectra acquired on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ (**a**) and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ (**b**) powders at 50 kHz MAS at a field $B_0 = 11.7$ T. The data have been fitted with a minimal number of Li sites: Li1, Li2 and Li3. Spinning sidebands of the three Li signals are indicated with asterisks. **c,** ^{19}F spin echo sum spectra acquired on as-synthesized $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ powders at 50 kHz MAS at a field $B_0 = 11.7$ T. The spectra are compared to the spin echo spectrum collected on LiF under similar conditions. Spinning sidebands of the sharp LiF-like signals are indicated with asterisks. Detailed explanations of the results are given in Methods section 'Supplementary Note 1'.
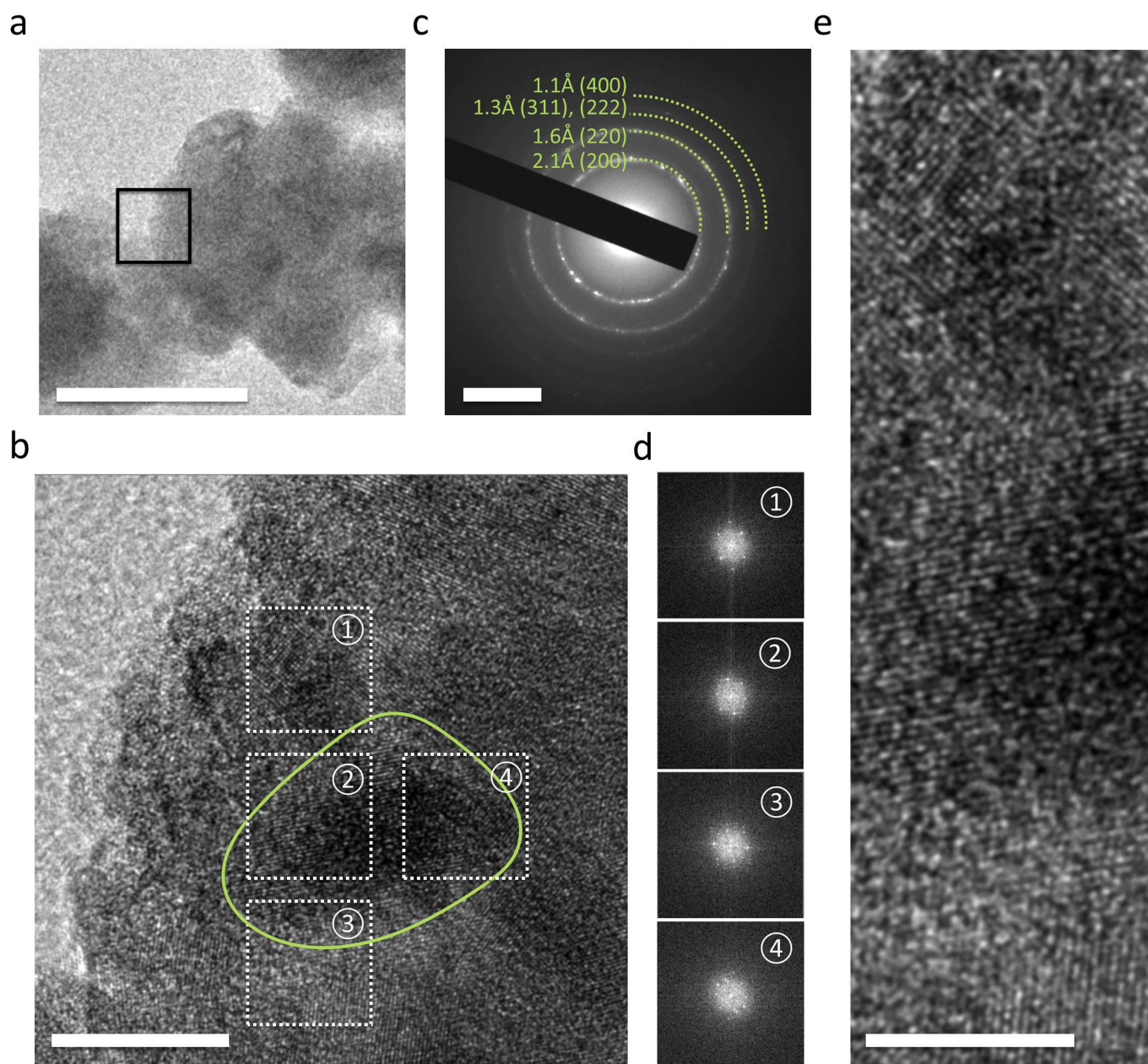
Extended Data Fig. 2 | Structural characterization of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F.
a, TEM image of as-synthesized Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F particles. Scale bar, 50 nm. **b**, A high-magnification TEM image of the area enclosed in a square in **a**. Scale bar, 10 nm. The yellow circle indicates the boundary of one of the many grains in the polycrystalline Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F particle. **c**, An electron diffraction pattern of the Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F particle.

Scale bar, 5 nm^{-1}. **d**, Fast Fourier-transformed (FFT) images of the dotted squared areas in **b**. **e**, The high magnification image across the squared areas 1, 2 and 3 in **b**. Scale bar, 5 nm. We can clearly observe lattice fringes and FFT peaks throughout the particle, indicating that our particles are made of small crystalline grains instead of amorphous phases.
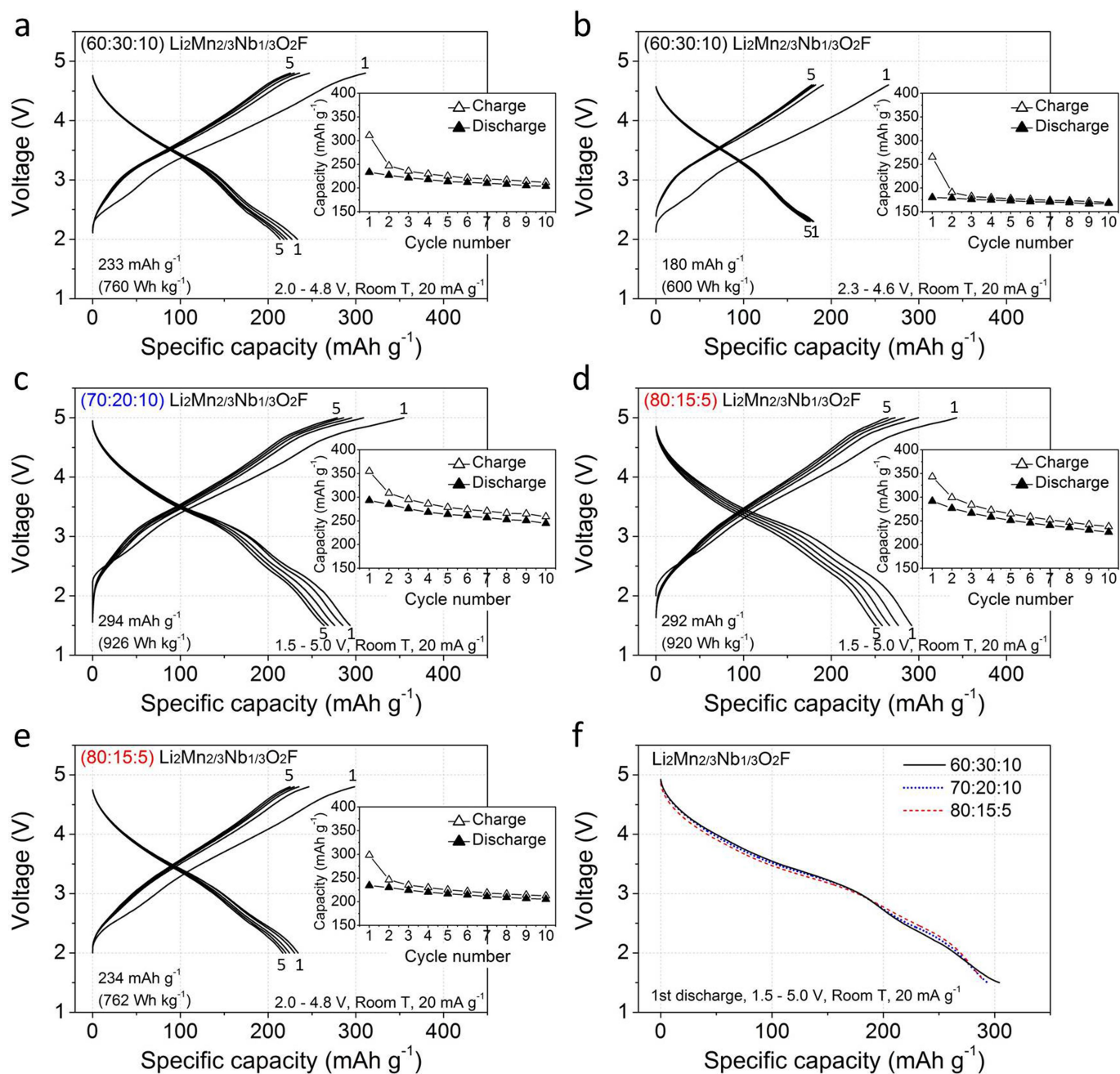
Extended Data Fig. 3 | Additional electrochemical data from Li₂Mn₂/₃Nb₁/₃O₂F. **a**, **b**, Voltage profiles of the 60:30:10 electrode (that is, 60 wt% Li₂Mn₂/₃Nb₁/₃O₂F: 30 wt% carbon black: 10 wt% PTFE) when cycled between 2.0 V and 4.8 V (**a**), and 2.3 V and 4.6 V (**b**) at 20 mA g⁻¹. **c**, **d**, Voltage profiles of the 70:20:10 (**c**) and the 80:15:5 (**d**) electrodes, when cycled between 1.5 V and 5.0 V at 20 mA g⁻¹. **e**, Voltage profiles of

the 80:15:5 electrode when cycled between 2.0 V and 4.8 V at 20 mA g⁻¹. **f**, A comparison of the first discharge profiles of the 60:30:10, 70:20:10 and 80:15:5 Li₂Mn₂/₃Nb₁/₃O₂F electrodes (1.5–5.0 V, 20 mA g⁻¹). The specific capacity was calculated on the amount of the Li₂Mn₂/₃Nb₁/₃O₂F powder in the cathode film.
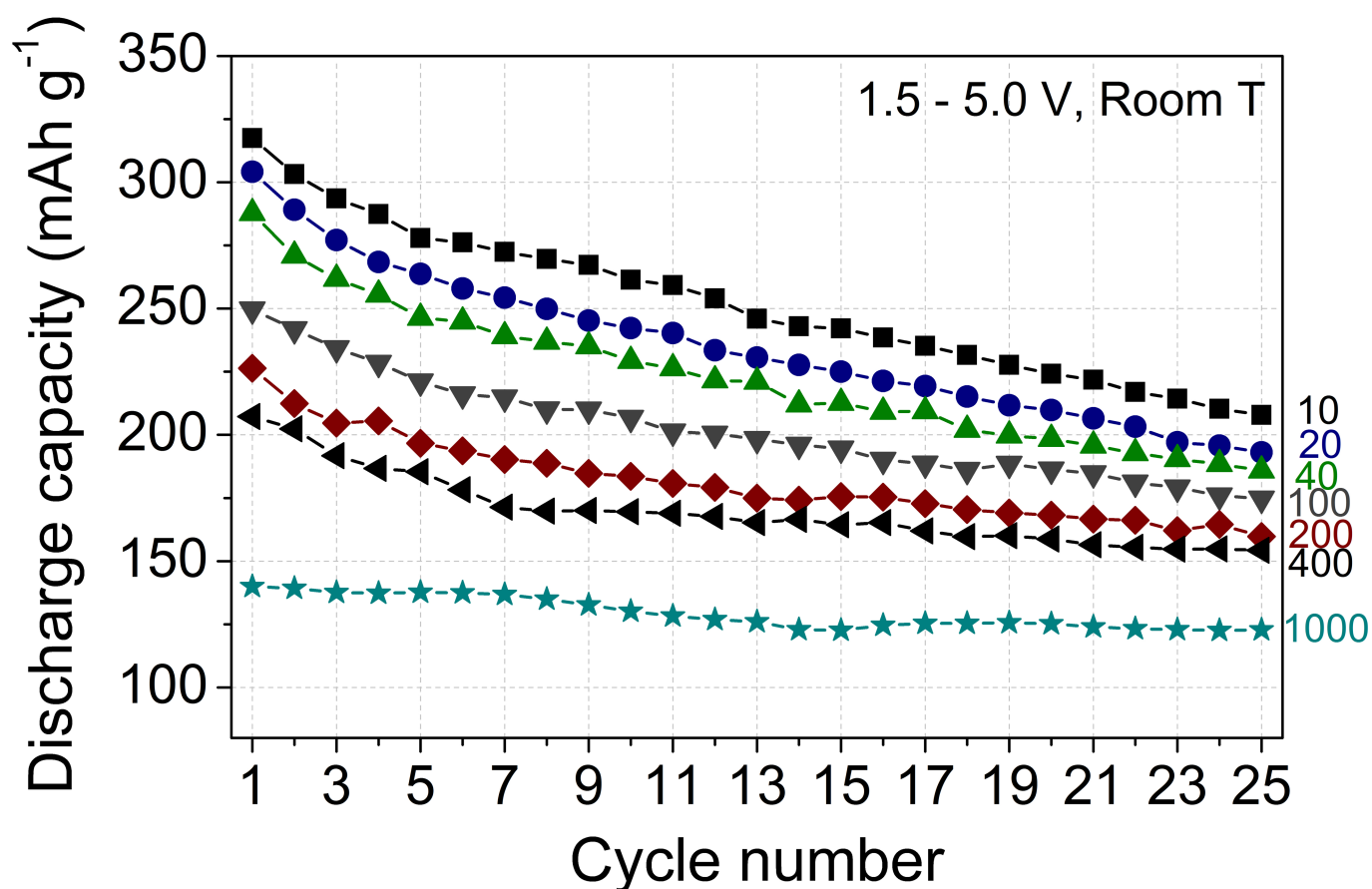
**Extended Data Fig. 4 | Discharge capacity retention.** The 60:30:10 $Li_2Mn_{2/3}Nb_{1/3}O_2F$: carbon black:PTFE electrode was cycled between 1.5 V and 5.0 V at room temperature at 10, 20, 40, 100, 200, 400 and 1,000 mA g^{-1}. A faster rate leads to less capacity fading during the initial 25 cycles. This is likely to be due to electrolyte decomposition per cycle occurring more (less) at a high voltage in a slower (faster) cycling test, which increases the impedance of a cell by creating a resistive surface layer and decreasing the ionic conductivity of the electrolyte.
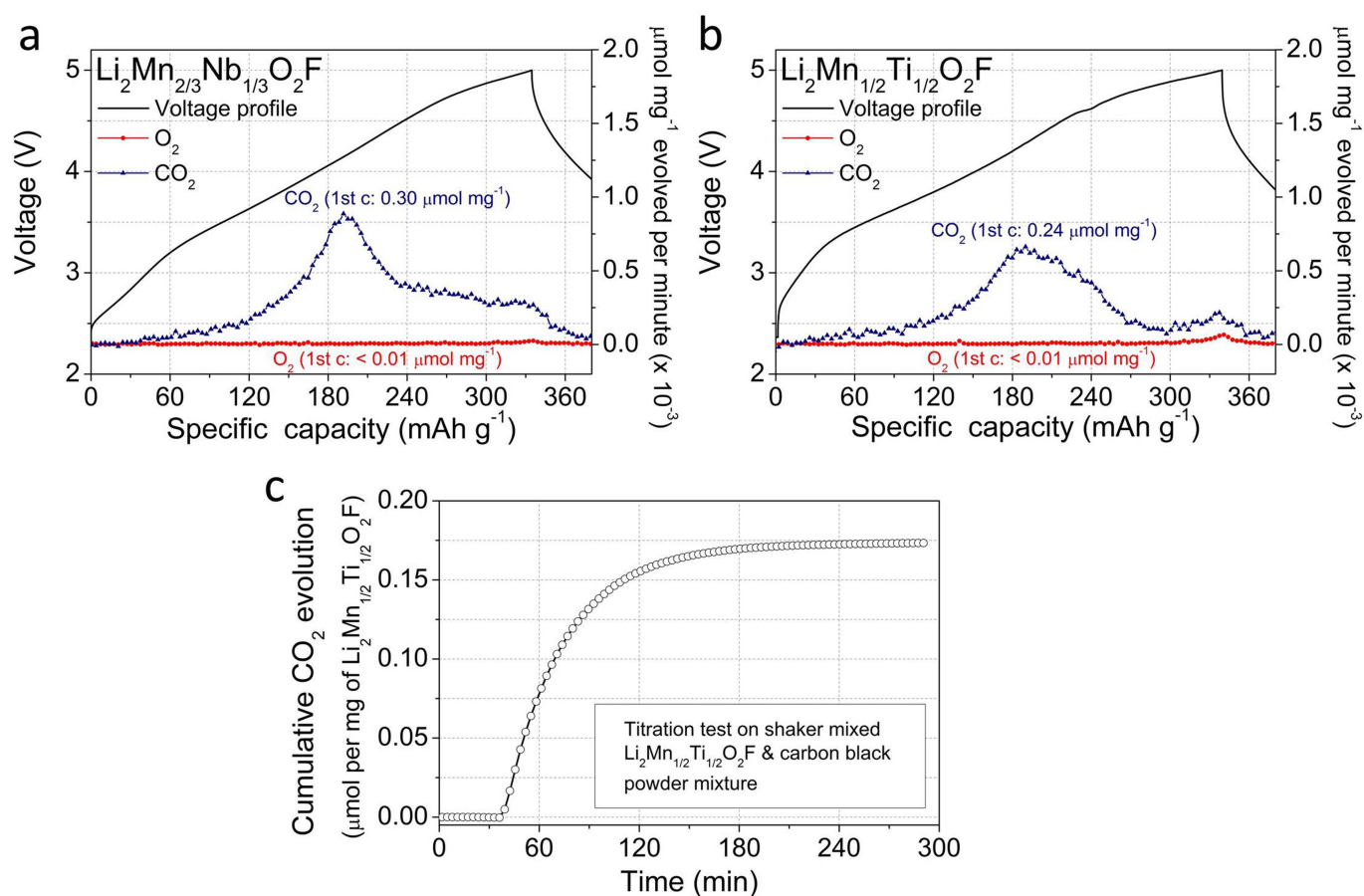
Extended Data Fig. 5 | Gas evolution measurements. a, **b**, Initial voltage profiles (black solid line) of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ (**a**) and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ (**b**), when charged to 5.0 V at a rate of 20 mA g^{-1}. DEMS results for O_2 (red circles) and CO_2 (blue triangles) evolution are also shown. **c**, Cumulative CO_2 evolution from shaker-mixed $Li_2Mn_{1/2}Ti_{1/2}O_2F$ and carbon black powder mixture, as a function of time during an acid titration test using 1 M H_2SO_4. Detailed explanations of the results are given in Methods section 'Supplementary Note 2'. 1st c, first charge.
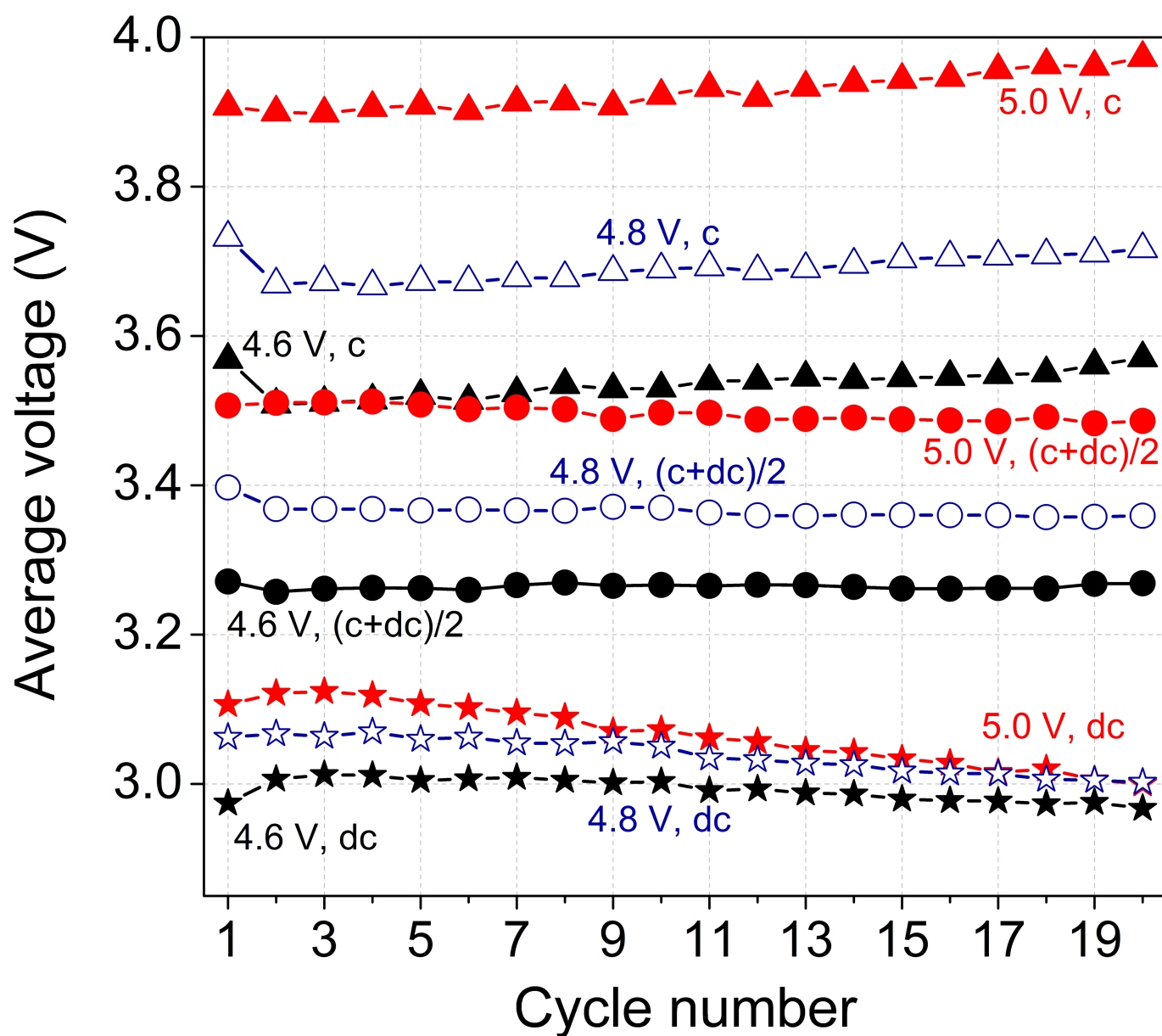
Extended Data Fig. 6 | Evolution of the charge and discharge voltages. Average charge voltage (triangles), discharge voltage (stars), and half of the charge–discharge voltage (circles) are shown when $Li_2Mn_{2/3}Nb_{1/3}O_2F$ is cycled between 1.5 V and 4.6 V, 1.5 V and 4.8 V, and 1.5 V and 5.0 V, at 20 mA g^{-1}. Detailed explanations of the results are given in Methods section 'Supplementary Note 3'. c, charge; dc, discharge.
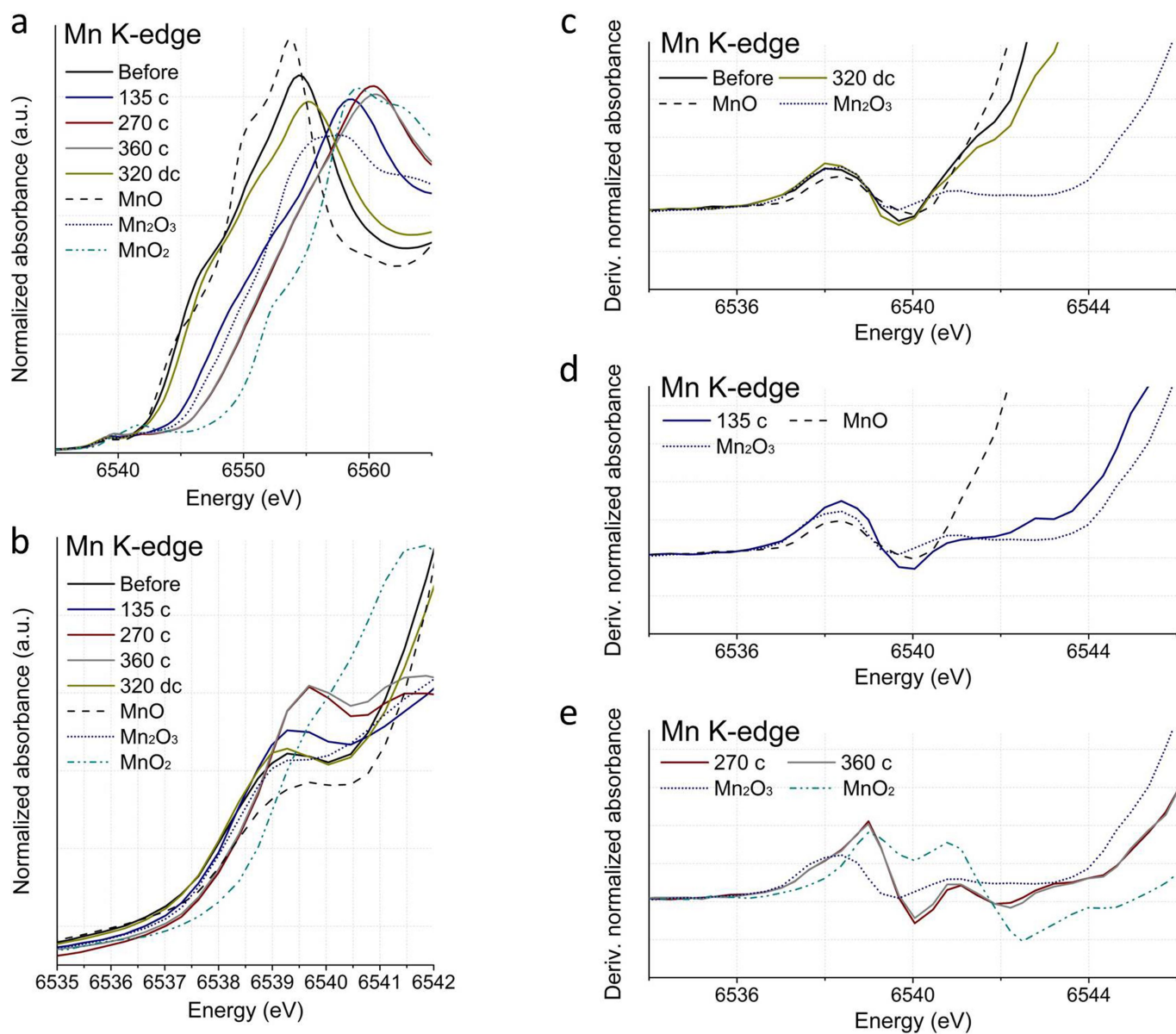
Extended Data Fig. 7 | XANES of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F. a, b, Manganese K-edge XANES spectra of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F: before cycle, after first charging to 135 mAh g^{-1}, 270 mAh g^{-1} and 360 mAh g^{-1}, and after first charging to 375 mAh g^{-1} then discharging to 320 mAh g^{-1}. **c–e,** First derivatives of normalized absorbance at the pre-edge region of Mn K-edge spectra of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F: **c,** before cycle and after first charging to 375 mAh g^{-1} then discharging to 320 mAh g^{-1}; **d,** after first charging to 135 mAh g^{-1}; and **e,** to 270 mAh g^{-1} and 360 mAh g^{-1}. Data from MnO, Mn$_2$O$_3$ and MnO$_2$ are presented for comparison. Detailed explanations of the results are given in Methods section 'Supplementary Note 4'.
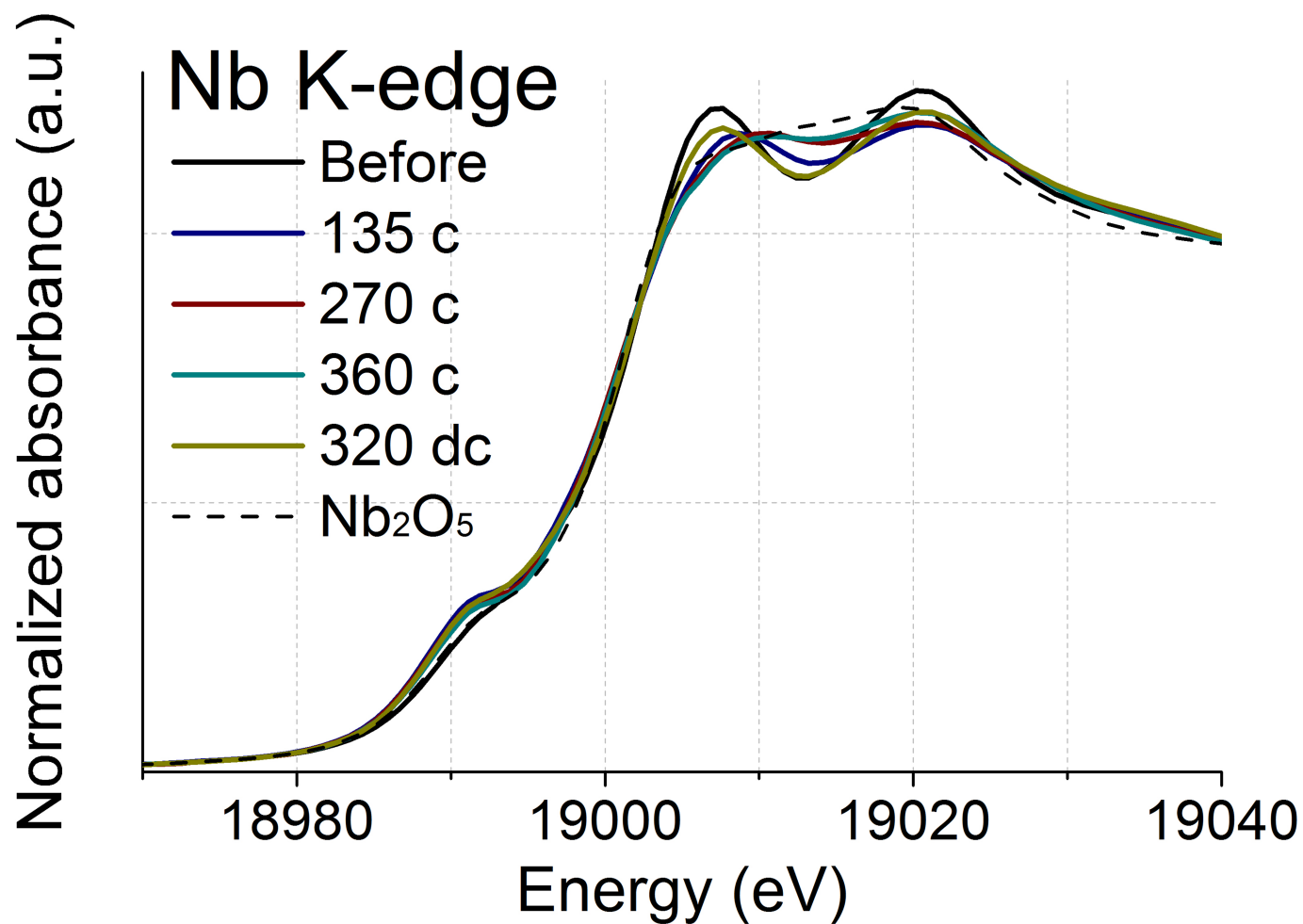
Extended Data Fig. 8 | Niobium K-edge XANES spectra of Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F obtained by hard XAS. Results are shown before cycle, after charging to 135 mAh g^{-1}, 270 mAh g^{-1} and 360 mAh g^{-1}, and after charging to 375 mAh g^{-1} then discharging to 320 mAh g^{-1}. The Nb K-edge XANES spectra of the Li$_2$Mn$_{2/3}$Nb$_{1/3}$O$_2$F samples are similar to that of Nb$_2$O$_5$ (Nb^{5+} reference), indicating that Nb in the compound stays as Nb^{5+} during cycling. The observable small shape changes are likely to be related to changes in local disorder and distortion[60].
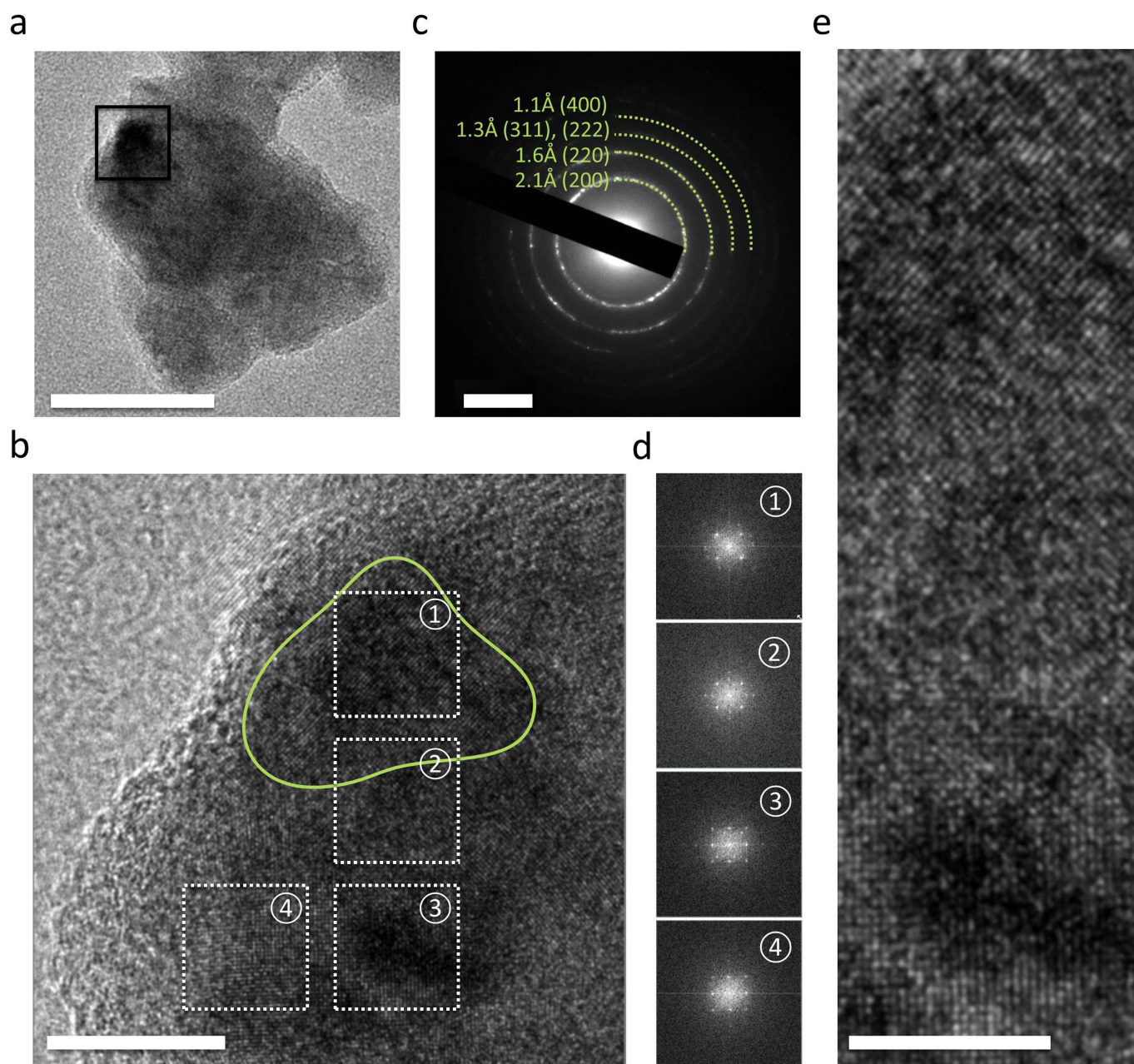
Extended Data Fig. 9 | Structural characterization of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F.
a, TEM image of as-synthesized Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F particles. Scale bar, 50 nm. **b**, A high-magnification TEM image of the area enclosed in a square in **a**. Scale bar, 10 nm. The yellow circle indicates the boundary of one of the many grains in the polycrystalline Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F particle. **c**, An electron diffraction pattern of the Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F particle.

Scale bar, 5 nm^{-1}. **d**, FFT images of the dotted squared areas in **b**. **e**, The high magnification image across the squared areas 1, 2 and 3 in **b**. Scale bar, 5 nm. We can clearly observe lattice fringes and FFT peaks throughout the particle, indicating that our particles are made of small crystalline grains instead of amorphous phases.
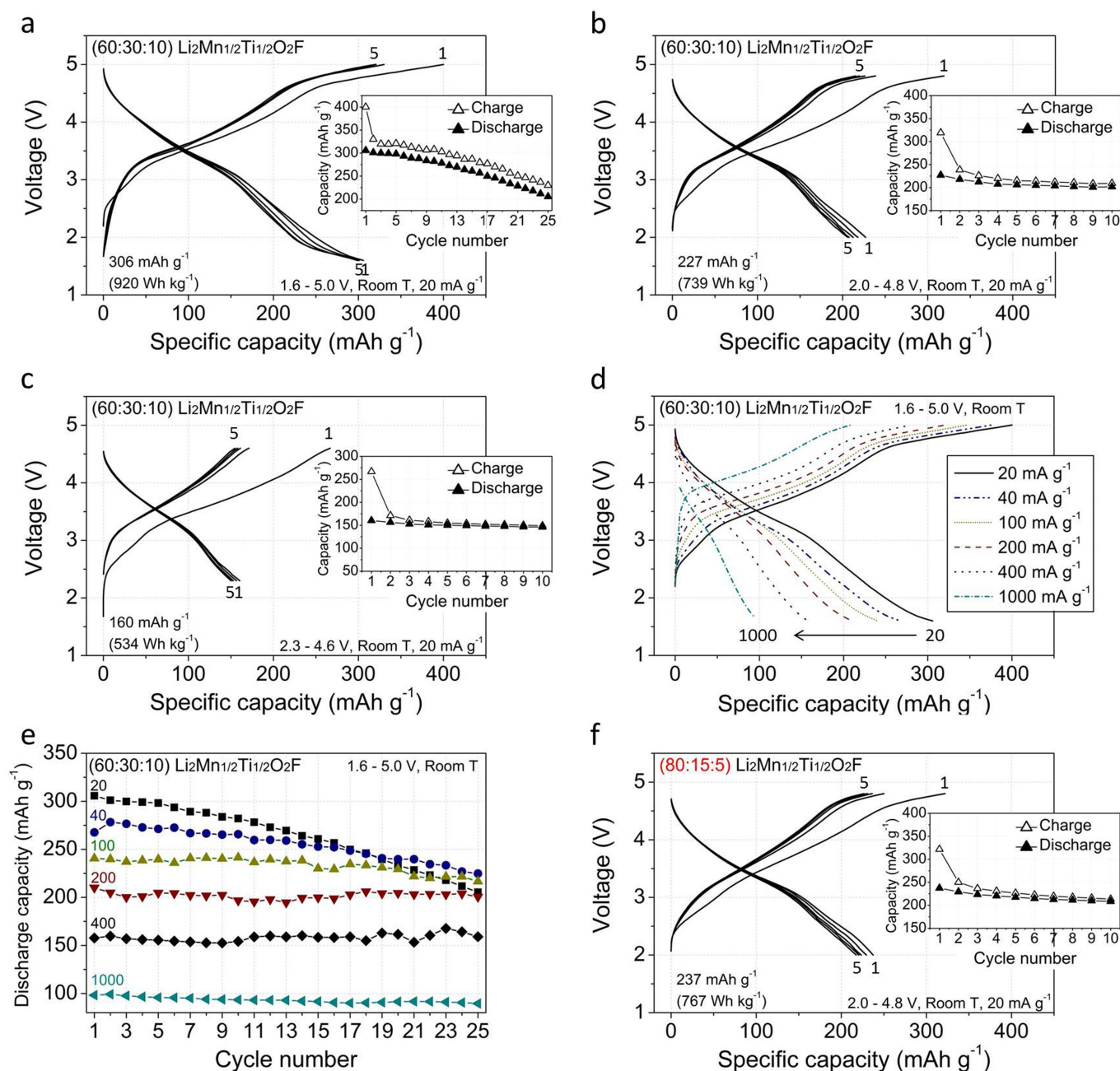
Extended Data Fig. 10 | Electrochemical properties of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F.
a–c, Voltage profiles and capacity retention of the 60:30:10
Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F:carbon black:PTFE electrode when cycled at 20 mA g^{-1}
at room temperature between 1.6 V and 5.0 V (**a**), 2.0 V and 4.8 V (**b**), and
2.3 V and 4.6 V (**c**). **d,** The initial charge–discharge profile of the 60:30:10
electrode when cycled between 1.6 V and 5.0 V at room temperature at

20, 40, 100, 200, 400 and 1,000 mA g^{-1}. **e,** The discharge capacities during
initial 25 cycles. **f,** Voltage profiles and capacity retention of the 80:15:5
electrode when cycled at 20 mA g^{-1} at room temperature between 2.0 V
and 4.8 V. The specific capacity was calculated on the amount of the
Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F powder in the cathode film. Detailed explanations of the
results are given in Methods section 'Supplementary Note 5'.
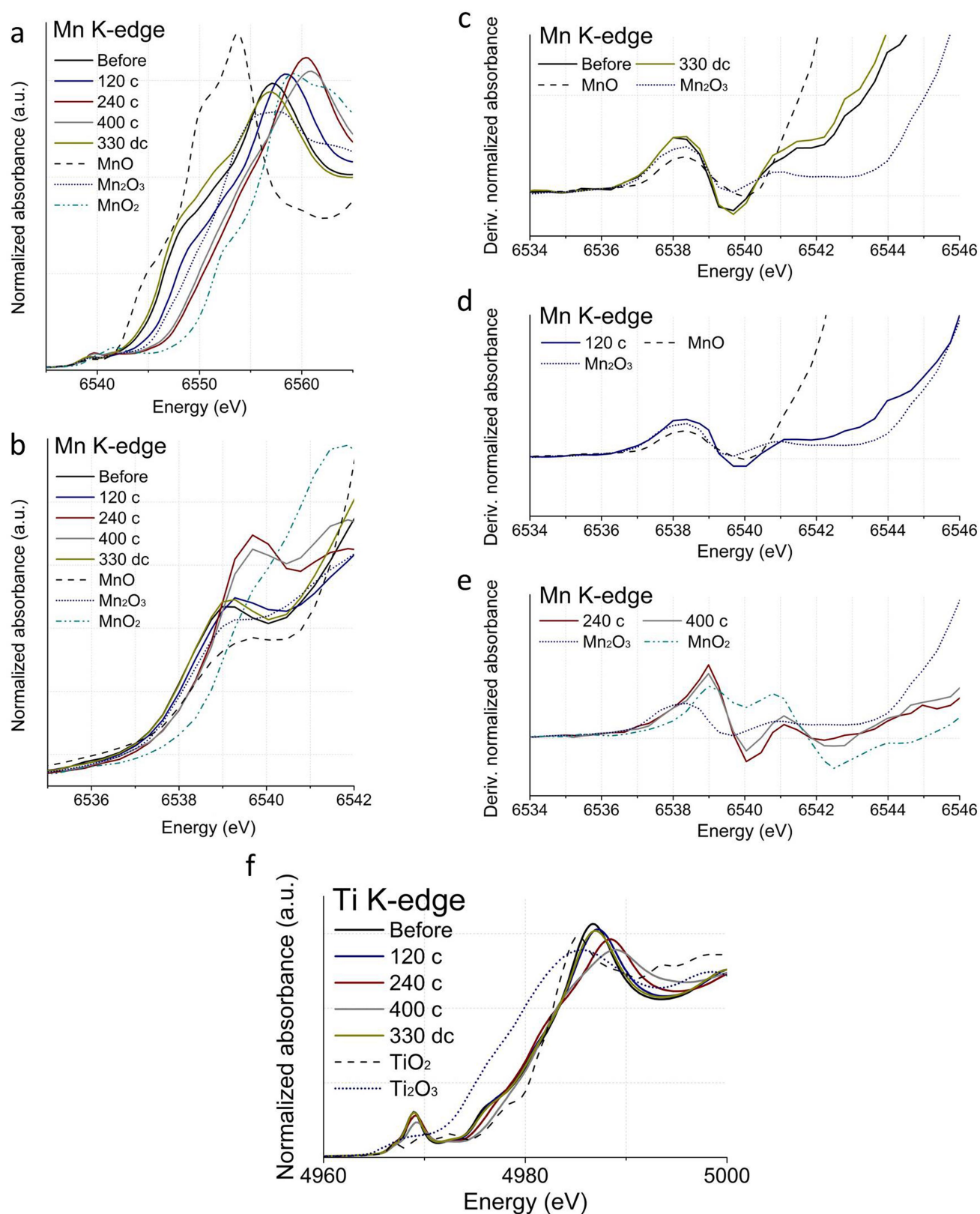
Extended Data Fig. 11 | XANES of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F. a, b, Manganese K-edge XANES spectra of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F: before cycle (black), 120 mAh g^{-1} charged (navy), 240 mAh g^{-1} charged (wine), 400 mAh g^{-1} charged (grey), 330 mAh g^{-1} discharged after a 400 mAh g^{-1} charge (dark yellow). **c–e**, First derivatives of normalized absorbance at the pre-edge region of Mn K-edge spectra of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F: **c**, before cycle and after first charging to 400 mAh g^{-1} then discharging to 330 mAh g^{-1}; **d**, after first charging to 120 mAh g^{-1}; and **e**, to 240 mAh g^{-1} and 400 mAh g^{-1}. **f**, Titanium K-edge XANES spectra of Li$_2$Mn$_{1/2}$Ti$_{1/2}$O$_2$F during the initial cycle. Data from MnO, Mn$_2$O$_3$, MnO$_2$, Ti$_2$O$_3$ and TiO$_2$ are presented for comparison. Detailed explanations of the results are given in Methods section 'Supplementary Note 6'.

Extended Data Table 1 | Structural parameters from the Rietveld refinements

Materials		$Li_2Mn_{2/3}Nb_{1/3}O_2F$	$Li_2Mn_{1/2}Ti_{1/2}O_2F$
Space group		Fm-3m	
R_{wp}		1.0023	1.1908
Goodness of Fit		0.7585	0.7494
B overall		0.1814 ± 0.0341	1.0236 ± 0.0297
Site 4a $(x, y, z) = (0, 0, 0)$	Li occupancy	0.6010 ± 0.0134	0.6537 ± 0.0135
	Mn occupancy	0.2244 ± 0.0033	0.1818 ± 0.0059
	Ti occupancy	0	0.1651 ± 0.0059
	Nb occupancy	0.1136 ± 0.0033	0
Site 4b $(x, y, z) = (0.5, 0.5, 0.5)$	O occupancy	0.6721 ± 0.0328	0.6442 ± 0.0384
	F occupancy	0.3279 ± 0.0328	0.3558 ± 0.0384
a (Å)		4.2615 ± 0.0008	4.2064 ± 0.0006
Volume (Å³)		77.39	74.43
Derived density (kg/l)		3.78	3.52

The Rietveld refinements are shown in Figs. 1b and 5a. The crystallographic information file of $Fm\overline{3}m$ $LiFeO_2$ (ICSD collection code 51208) was used as an input file. A pseudo-Voigt fit was used ($U,V,W = 8.0691, -0.9697, 1.3778$ for $Li_2Mn_{2/3}Nb_{1/3}O_2F$, and $5.8736, -1, 1.4118$ for $Li_2Mn_{1/2}Ti_{1/2}O_2F$). The atomic occupancies were initially set to the atomic ratio obtained from elemental analysis by direct-current plasma emission spectroscopy and an ion-selective electrode, based on which the lattice parameters were first refined. We then further refined the lattice parameters and the atomic occupancies together. Transition-metal occupancies were first refined freely. Then O and F occupancies were individually refined with a constraint of their occupancies summing to 1. Finally, all atomic occupancies including Li occupancy were simultaneously refined with the additional constraint that the total transition-metal occupancy should stay unchanged during this final process. However, as O and F are difficult to distinguish by XRD, and Li cannot be seen clearly, their occupancy values are more subject to error.

Extended Data Table 2 | Target versus measured atomic ratio of $Li_2Mn_{2/3}Nb_{1/3}O_2F$ and $Li_2Mn_{1/2}Ti_{1/2}O_2F$ compounds

Materials	$Li_2Mn_{2/3}Nb_{1/3}O_2F$ (Li : Mn : Nb : F)	$Li_2Mn_{1/2}Ti_{1/2}O_2F$ (Li : Mn : Ti : F)
Target atomic ratio	2 : 0.666 : 0.333 : 1	2 : 0.5 : 0.5 : 1
Measured atomic ratio	1.852 : 0.660 : 0.333 : 1.05	2.01 : 0.514 : 0.475 : 1.05

Measurements were made by direct-current plasma emission spectroscopy (Li, Mn, Nb, Ti) and with an ion-selective electrode (F).

ARTICLE

Observed fingerprint of a weakening Atlantic Ocean overturning circulation

L. Caesar[1,2]*, S. Rahmstorf[1,2]*, A. Robinson[1,3,4,5], G. Feulner[1] & V. Saba[6]

The Atlantic meridional overturning circulation (AMOC)—a system of ocean currents in the North Atlantic—has a major impact on climate, yet its evolution during the industrial era is poorly known owing to a lack of direct current measurements. Here we provide evidence for a weakening of the AMOC by about 3 ± 1 sverdrups (around 15 per cent) since the mid-twentieth century. This weakening is revealed by a characteristic spatial and seasonal sea-surface temperature 'fingerprint'—consisting of a pattern of cooling in the subpolar Atlantic Ocean and warming in the Gulf Stream region—and is calibrated through an ensemble of model simulations from the CMIP5 project. We find this fingerprint both in a high-resolution climate model in response to increasing atmospheric carbon dioxide concentrations, and in the temperature trends observed since the late nineteenth century. The pattern can be explained by a slowdown in the AMOC and reduced northward heat transport, as well as an associated northward shift of the Gulf Stream. Comparisons with recent direct measurements from the RAPID project and several other studies provide a consistent depiction of record-low AMOC values in recent years.

The AMOC is one of Earth's major ocean circulation systems, redistributing heat on our planet and thereby affecting its climate. At the same time, it is a highly nonlinear system with a critical threshold, depending on a delicate balance of temperature and salinity effects on density, and is considered one of the main tipping elements of the Earth system[1,2]. Changes in Atlantic overturning have been responsible for some of the strongest and most rapid climate shifts during the Quaternary Period (the past 2.6 million years)[3]. These historical changes in the AMOC have not only affected the North Atlantic and surrounding landmasses, but have also had global impacts. For example, a slowdown of the AMOC is associated with a southward shift of the tropical rainfall belt and a warming of the Southern Ocean and Antarctica (the 'see-saw' response)[2,3].

Given the potentially disruptive impact of a major change in the AMOC, it is imperative to better understand whether and how the AMOC is responding to modern anthropogenic warming. Direct continuous measurements of the AMOC have only been available for a little over a decade and are therefore probably dominated by natural variability[4]. The longer-term evolution of the AMOC needs to be reconstructed from indirect indicators. Based on the observed cooling trend in the subpolar Atlantic since the early twentieth century, recent studies have suggested that the AMOC may have slowed over this period[5–7]. However, it has also been suggested that another mechanism could explain the subpolar Atlantic cooling, for example, the increasing aerosol load of the atmosphere[8].

Here we use the latest high-resolution climate model results to identify a characteristic sea-surface temperature (SST) fingerprint, consisting of a cooling in the subpolar gyre region and a warming in the Gulf Stream region, which in the climate model is associated with an AMOC reduction in response to rising atmospheric carbon dioxide (CO_2) levels[9]. We then compare this fingerprint with the observed SST evolution since the late nineteenth century, including consideration of the seasonal cycle. We use the climate-model ensemble of the Coupled Model Intercomparison Project Phase 5 (CMIP5) to test and calibrate

a revised AMOC index, and we present a new reconstruction of the AMOC evolution for the period 1870 to 2016. This index reaches record-low values in the past few years and, for the periods of overlap, is consistent with direct measurements, reanalysis data of the AMOC since 1995 and other AMOC studies.

Comparing climate model and SST observations

We use the CM2.6 coupled global climate model, which provides high horizontal resolution of around 50 km in the atmosphere and 10 km in the ocean (see Methods). The latter is important for analysing SST data because high resolution helps to reduce regional SST biases[10]. The model resolves mesoscale ocean eddies[11] and shows a more realistic simulation of the Gulf Stream relative to coarser model versions. In particular, this model practically eliminates a bias in the separation point of the Gulf Stream from the United States' coastline (leading to a warm and salty bias along the continental shelf), which is common in coarser climate models assessed by the Intergovernmental Panel on Climate Change (IPCC)[9]. After appropriate spin-up, we used two simulations: a control simulation of 80 years' duration with CO_2 concentrations fixed at the 1860 level, and a run in which atmospheric CO_2 increased by 1% per year over 70 years until it doubled, and then remained at this level for another 10 years.

Figure 1 shows the linear trend in SST over the 'CO_2-doubling' experiment and the corresponding control run, compared with the observed trend from 1870 to 2016 (owing to the extreme computational costs of the CM2.6 model, neither a simulation with historic forcing nor ensemble studies are available). The trend pattern of the observed SSTs is not sensitive to the choice of the time interval used to calculate the linear trend (see Extended Data Fig. 1). Figure 1 shows that the control run is almost free of SST trends, and that the observed SST trend pattern resembles that measured in the CO_2-doubling experiment. To account for the much larger global SST warming (by a factor of four) seen in the model experiment compared with observations, in Fig. 2 we divide both patterns by the global mean SST trend

CM2.6 2×CO$_2$ CM2.6 control   HadISST data

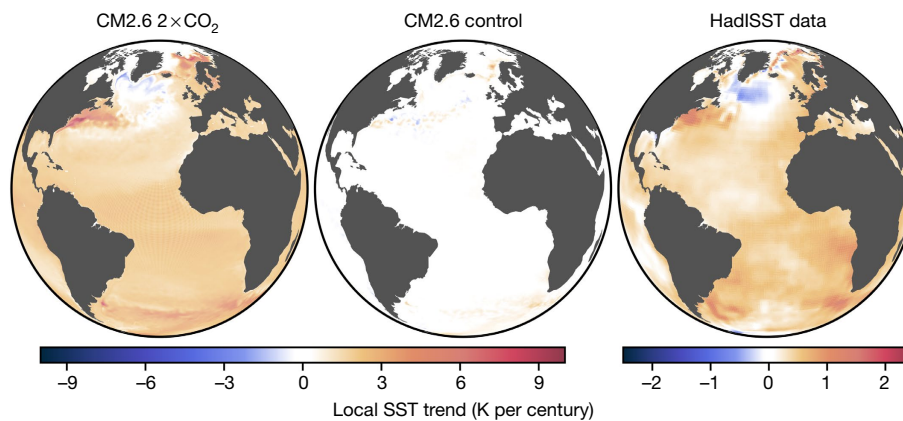Local SST trend (K per century)

Fig. 1 | Comparison of SST trends in model and observations. Left and middle, linear SST trends obtained using the CM2.6 climate model of the Geophysical Fluid Dynamics Laboratory (GFDL) during a CO$_2$-doubling experiment (left) and in a control run with fixed CO$_2$ concentrations (middle). Right, observed SST trends from 1870 to 2016 (HadISST data). We used data from the November–May season. Note the different scales related to the differing amounts of CO$_2$ forcing between model and observations.

to normalize the amplitude. A global view of these SST trends is shown in Extended Data Fig. 2.

The comparison of the normalized modelled and observed SST trend patterns (Fig. 2) shows a remarkable resemblance, especially when focusing on the northern Atlantic—the area where SSTs are most affected by changes in the AMOC. Both patterns comprise an area of below-average warming (normalized trend < 1) and cooling (normalized trend < 0) in the subpolar gyre region. This lack of warming or cooling is associated with a slowdown of the AMOC by around 4 sverdrups (Sv; 1 Sv = 10^6 m^3 s^{-1})—as predicted by the CM2.6 simulation (see Fig. 3)—and a corresponding reduction in heat transport into that region. This feature is accompanied by an above-average warming (normalized trend > 1) in the vicinity of the Gulf Stream, which is enhanced by up to a factor of four–five over the global mean warming (for a definition of the regions, see inset of Fig. 3). The median trend of the subpolar gyre region is located at the third percentile of all trends in the observational data, and at the first percentile in the model. The median trends in the Gulf Stream region are located at the 96th and 98th percentiles of all trends in the observational data and model, respectively (see Methods and Extended Data Fig. 3). We define the combination of these features as the AMOC fingerprint, as both signals can be physically linked to changes in the AMOC.

Although the cold patch in the subpolar gyre region has previously been connected to a slowdown of the AMOC[7] and is present in the CMIP5 simulations[12], here we are able to link the extreme warming observed along the US northeast coast to the Gulf Stream shifting northwards and closer to shore as a consequence of an AMOC slowdown (see Extended Data Fig. 4a). An opposite (that is, southward) Gulf Stream shift has previously been found as a response to an AMOC strengthening in idealized model simulations in which the AMOC was deliberately enhanced by an imposed density anomaly in the deep overflow from the Nordic Seas; this overflow feeds the lower branch of the AMOC[13], the deep western boundary current (DWBC). The physical mechanism of the interaction of the DWBC with the Gulf Stream at their crossing point is a robust mechanism that is known from theory and from both conceptual and more complex models: it is a consequence of vorticity conservation on a rotating sphere[14]. The downslope flow of the DWBC in the crossover region leads to vortex stretching, which must be balanced higher up in the water column, leading to the formation of a northern recirculation gyre that forces the Gulf Stream to separate from the US east coast. As the flow of the DWBC is strengthened, the recirculation gyre becomes stronger and the separation point of the Gulf Stream moves southwards. Given that the Gulf Stream transports warm water, this signal is reflected in the SST. For a more detailed discussion of this mechanism, see Methods.

The physical mechanism behind the warming also explains why it cannot be seen in climate models with a coarser ocean resolution,

including versions that are similar to CM2.6, with the same atmosphere but a coarser ocean resolution. Only the high-resolution model accurately represents the formation of the northern recirculation gyre and thus the correct coastal separation position of the Gulf Stream, which is a necessary condition for modelling the shifts in the Gulf Stream that are due to changes in AMOC strength. The northward shift of the warm water of the Gulf Stream leads to extreme warming along the US coast and a cooling to the south of this warming (as can be seen by the blue area to the south of the Gulf Stream in the CM2.6 simulation; Fig. 2). Another indication of a northward shift of the Gulf Stream in the CM2.6 model is enhanced warming of ocean-bottom temperatures on the continental shelf, particularly in the Gulf of Maine, as a result of a poleward retreat of the Labrador Current following the northward shift[9]. This warm part of the AMOC fingerprint cannot be explained by aerosol shading. The cooling in the subpolar gyre region in the CM2.6 model cannot be caused by aerosols either, because the modelled response is entirely CO$_2$-driven—that is, no aerosol forcing was prescribed. This strongly supports earlier arguments against the aerosol hypothesis[15].

We have looked for the fingerprint of an AMOC slowdown in seven available observational SST data products (Extended Data Fig. 5). All of these datasets show the cold patch in the subpolar Atlantic, and, to a greater or lesser extent, the enhanced warming inshore of the Gulf Stream. The weaker cooling signal just south of this warming cannot be seen in most of the observational datasets (except the COBE data; see Extended Data Fig. 5). This could be because of the lower spatial resolution of the observational data products and the smaller AMOC decline in the observations as compared with the model simulation. The data products are distinct partly because of the different input databases used, and because of different degrees of data homogenization, bias adjustment, averaging and interpolation, which preserve different amounts of spatial and temporal structure (see Extended Data Table 1). The main difference is that, for example, the ERSST data concentrate on the preservation of temporal structure, whereas the HadISST data focus on the preservation of spatial structure. As we are interested in the spatial pattern of longer-term trends, in Fig. 2 we show the SST data with the best combination of spatial resolution (1.0 × 1.0 degrees), spatial preservation and quality control, namely, the HadISST data[16].

We note that the sea-ice-covered regions of the Arctic Ocean show no temperature trend, consistent with the assumption that SST remains close to freezing point there. In the observations, this blue area is crossed by a red line where the sea-ice margin has retreated (Fig. 2). The linkages of the AMOC in the open Atlantic to the northward flow of Atlantic waters past Iceland warrant further investigation, but are beyond the scope of this paper.

Finally, both model and data show widespread above-average warming in the South Atlantic, consistent with the temperature see-saw effect of an AMOC decline leading to reduced northward ocean heat
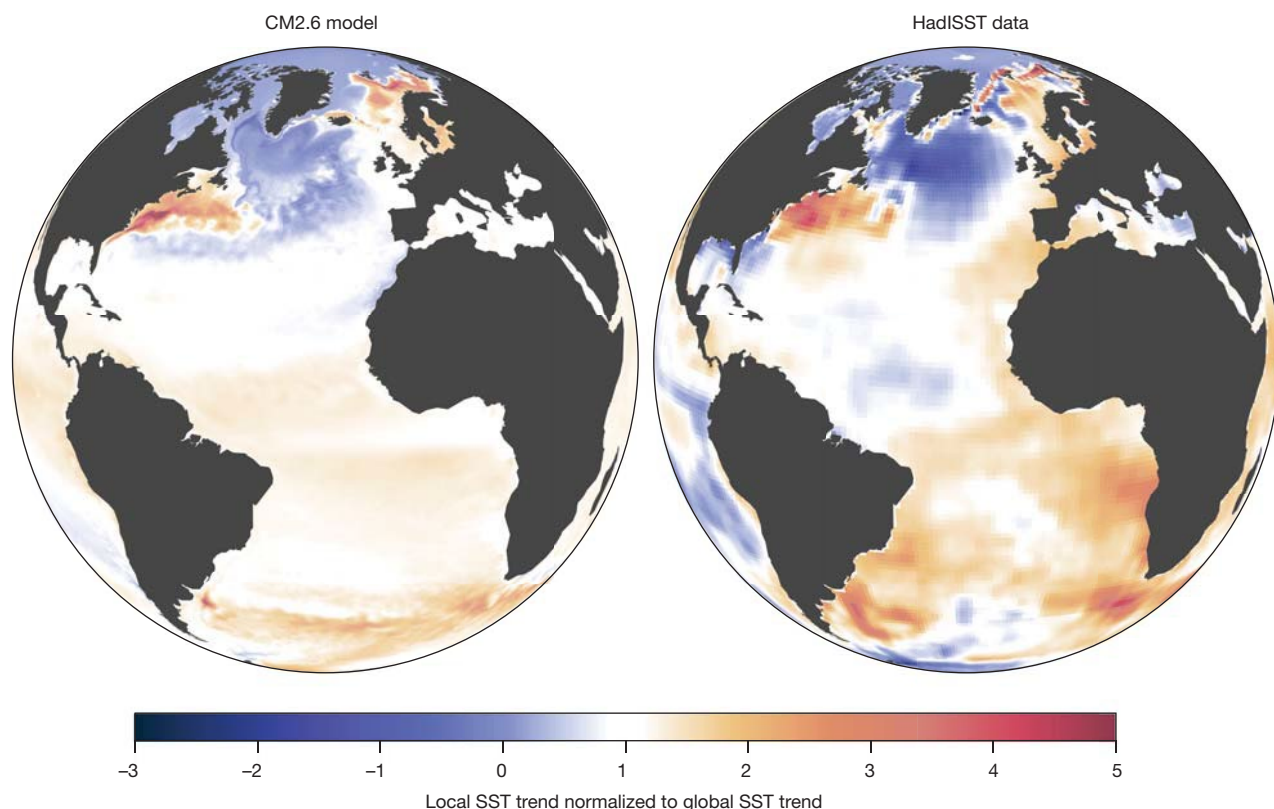
CM2.6 model HadISST data

Local SST trend normalized to global SST trend

Fig. 2 | Comparison of normalized SST trends. Left, linear SST trends during a CO$_2$-doubling experiment using the GFDL CM2.6 climate model. Right, observed trends during 1870–2016 (HadISST data). Both sets of data are normalized with the respective global mean SST trends, and in both cases we used data from the November–May season. Regions that show cooling or below-average warming are shown in blue; regions that show above-average warming are in red. Owing to the much greater climate change in the CO$_2$-doubling experiment, the signal-to-noise ratio for the modelled SST trends is better than that for the observations.

transport across the equator[17,18]. The observations show particularly strong warming along the Benguela Current and its northward extension towards the Gulf of Guinea. This is a common response in climate models to an AMOC weakening[2,19,20], and is related to a reduced cold northward flow, but is not seen in the CM2.6 simulations. This omission might be related to the model's representation of the AMOC or of wind-driven circulation in the South Atlantic, and needs further investigation.
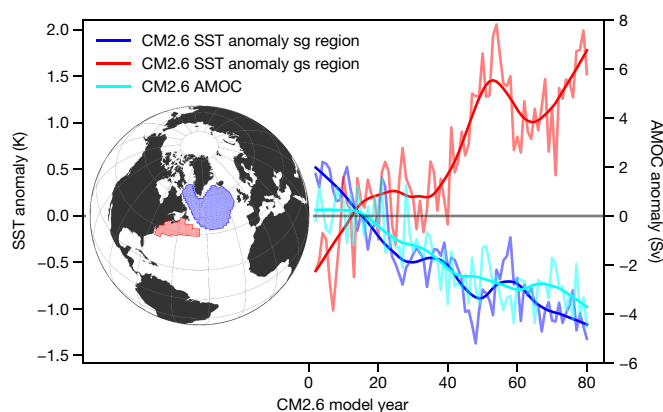

Fig. 3 | Comparison of time series of SST anomalies and the strength of the overturning circulation in the CM2.6 model. The graph shows time series of SST anomalies (relative to global mean SSTs) in the subpolar gyre (sg; dark blue) and Gulf Stream (gs; red) regions in the CO$_2$-doubling run relative to the control run, as predicted by the CM2.6 model. These two regions are defined as shown in the inset (see Methods). The anomaly of the actual AMOC overturning rate relative to the control run is also shown (light blue). Thin lines show individual years (November to May for SSTs), and thick lines show 20-year locally weighted scatterplot smoothing (LOWESS) filtered data. Using the CMIP5 ensemble, we independently determined a conversion factor of 3.8 Sv K^{-1} between the SST anomaly and the AMOC anomaly.
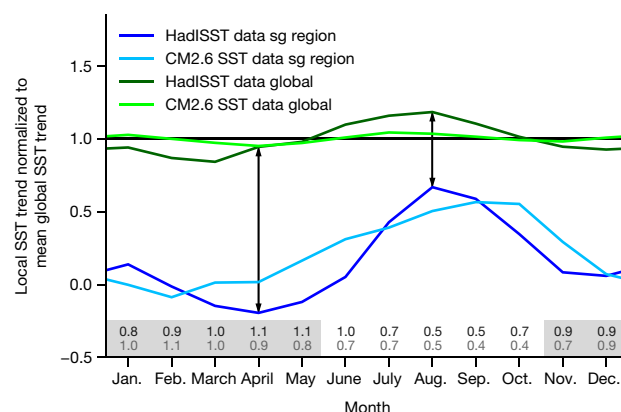
Fig. 4 | Seasonal variation in SSTs in the subpolar gyre region. We show here the seasonal cycle in the normalized SST trend in the subpolar gyre (sg) region for the CM2.6 model (light blue) and HadISST data (dark blue). A value of 1 represents annual-mean, global-mean warming. In addition, we show the seasonal cycle of the normalized global-mean SST trend for the model (light green) and observations (dark green). The SST trends in the subpolar gyre region are well below the global-mean warming year-round (differences are given in numbers along the x axis for the CM2.6 model (light grey) and the HadISST data (dark grey) and highlighted by arrows), yet are smallest during the cold part of the year for both observations and model.

Fig. 5 | Results of the CMIP5 ensemble regression analysis. The graph shows the linear trend in the simulated AMOC decline versus the SST-based AMOC index (November–May data) in 'historic' climate model runs from 1870 to 2016, using the CMIP5 climate model ensemble. (The runs were extended from 2006 to 2016 with simulations of the RCP8.5 scenario.) Orthogonal regression analysis was performed with $n = 12$ models (indicated by coloured symbols). The grey area marks the 2σ confidence 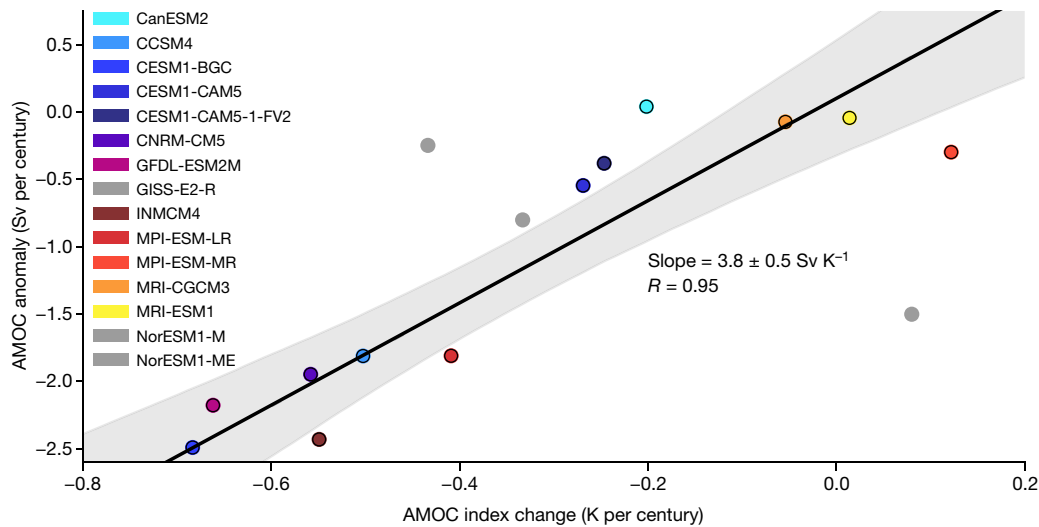interval. The three models labelled in grey were not included in the regression owing to unrealistic AMOC representation; see Methods.

The subpolar cold patch as an AMOC indicator

The surface temperature in the subpolar gyre region, relative to the large-scale temperature trend, has been proposed as an index for longer-term AMOC variations[7]. Here we test and develop this concept further. Figure 4 compares the seasonal cycle in the linear SST trend in the subpolar gyre region from the HadISST data since 1870 with the 80-year CO_2-doubling experiment. The figure shows that the cooling (relative to the global mean SST) in this region is most pronounced during winter and spring. This is to be expected if the relative cold in this area is due to an AMOC slowdown and therefore driven by the ocean. In summer, a shallow surface mixed layer develops that is more susceptible to surface forcing than to horizontal heat advection, so the cold patch can be effectively capped and hidden by a warm surface layer. It typically re-emerges in autumn.

Given this result, in Fig. 2 we show the linear trends for November to May and below we propose an improved AMOC index based on these months, with a better signal-to-noise ratio than that obtained using annual data. The AMOC fingerprint pattern itself is not sensitive to the choice of the winter and spring seasons, as the linear trends of the annual data show (Extended Data Fig. 1).

Performance of the AMOC index in models

Given the hypothesis that a slowdown of the AMOC leads to a region of relative cooling near the subpolar gyre and a region of above-average warming in the vicinity of the Gulf Stream, we test whether in the models the temperatures in these regions can be used to reconstruct changes in the AMOC.

Figure 3 shows time series of the mean temperatures of the subpolar gyre (sg, dark blue line) and the Gulf Stream (gs, red line) regions relative to—that is, minus—the global mean SST. The averaging regions are defined as shown in the inset of Fig. 3 (see Methods).

The two modelled SST time series are anti-correlated ($R = -0.73$), yet the pronounced temperature maximum in the Gulf Stream region around model year 50 (red line), which is unrelated to an AMOC change in the model (light blue line), suggests that variability due to factors other than the AMOC is substantially affecting the temperature of the warm patch. This is to be expected particularly for the coastal waters in the Gulf Stream region, which are more susceptible to wind-forced SST changes—for example, owing to the presence of strong horizontal gradients and coastal upwelling or downwelling. In accordance with this, the observed time series for the warm and cold patches are only moderately anti-correlated ($R = -0.36$). This variability, unrelated to the AMOC, makes the warm patch unsuitable for use as an AMOC proxy owing to its poor signal-to-noise ratio, in contrast to the subpolar cold patch (see below). To maximize the signal-to-noise ratio, we base the AMOC index definition only on the subpolar gyre data (see Methods).

To test the ability of this index of detecting past AMOC changes, we turn to the CMIP5 coupled climate model ensemble[20], using all simulations for which an AMOC diagnostic is available ($n = 15$; Extended Data Table 1). The region defining the subpolar cold patch is chosen to be large enough to encompass the cooling found across all models, because its exact location differs in each model. Figure 5 shows the linear 1870–2016 trend in the AMOC index, as well as in the actual AMOC, in these models. The correlation for the models with a realistic AMOC has $R = 0.95$, so the AMOC variation explains 89% of the
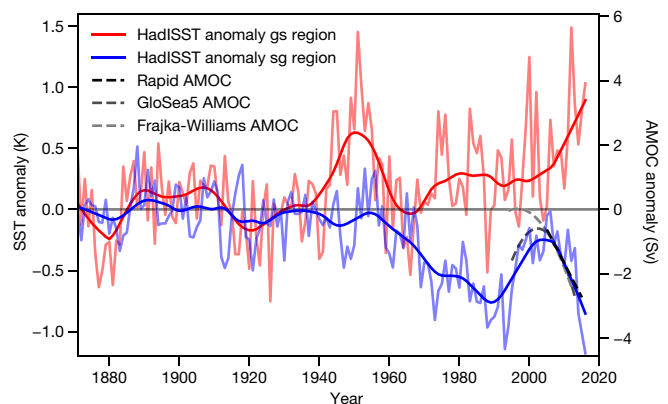
Fig. 6 | Comparison of time series of SST anomalies and the strength of the overturning circulation in observations. Shown are time series of SST anomalies with respect to the global mean SST in the subpolar gyre (sg) and the Gulf Stream (gs) regions (HadISST data). The graph also includes the trend of in situ AMOC monitoring by the RAPID project[21], an ocean reanalysis product (GloSea5[22]) and a reconstruction from satellite altimetry and cable measurements[23]. Thin lines show individual years (November–May for SSTs) and thick lines show smoothed data (20-year LOWESS filtering for the SST data and quadratic/linear fits for the AMOC data).

variance in the AMOC index. This confirms that the AMOC (at least on this long timescale) is indeed the dominant factor controlling the SST anomaly in the subpolar Atlantic. Hence the AMOC index can be used with confidence to identify the AMOC decline since 1870. The total-least-squares line shown in Fig. 5 has a slope of $3.8 \, \text{Sv} \, \text{K}^{-1}$ and an intercept of 0.1 Sv for the chosen subpolar gyre region (for more information on the regression, see Methods). The very small intercept value suggests that factors other than the AMOC have a minor influence on SST changes in the subpolar Atlantic. For example, a local aerosol cooling effect, relative to the global mean SST change, would cause a systematic offset in this regression. Given that this offset is negligible, however, the slope value of $3.8 \, \text{Sv} \, \text{K}^{-1}$ can be used to calibrate between the AMOC index and the AMOC strength.

AMOC time evolution

In Fig. 6 we show the time evolution of the AMOC, reconstructed from observational SST data (blue curve) from the period 1870–2016 using the calibration factor $3.8 \, \text{Sv} \, \text{K}^{-1}$ found from the CMIP5 models (for a comparison with the earlier AMOC index[7], see Extended Data Fig. 6). This time evolution suggests that the AMOC reached a minimum around 1990, recovered to a peak value in the early 2000s, and then declined again. As shown, this time evolution is consistent with the linear decline measured by the RAPID project (at 26° N)[21] since 2004, with that reconstructed by the GloSea5 ocean reanalysis[22] since 1995, and with a reconstruction from satellite altimetry and cable measurements[23]. It is also consistent with the finding[24] of a reduction in AMOC strength of approximately 2.6 Sv from the end of the 1950s until today, and with the observation[25] of an AMOC strengthening from the 1980s until the mid-2000s. An analysis of recent (2004–2016) subsurface temperature data[26] found cold subsurface anomalies around the latitude of the Gulf Stream (38° N) that could be associated with a shift in the meridional position of the Gulf Stream towards the north, supporting our argument for such a shift in response to an AMOC decline.

The observed index decline of $-0.44 \, \text{K}$ per century translates into an AMOC trend of $-1.7 \, \text{Sv}$ per century, or a 2.3-Sv linear weakening over the 136-year period. As Fig. 5 shows, this AMOC decline is within the range of AMOC decline predicted by the CMIP5 climate models in response to historic (mostly anthropogenic) forcing. Considering the 20-year smoothed curve rather than the linear trend, the AMOC weakening until today has been around 3 Sv, and has mainly occurred since the 1950s (Fig. 6).

Comparing the SST anomalies in the CM2.6 model (Fig. 3) and observations (Fig. 6), one can see that generally they show similar magnitudes of interannual and interdecadal variability. To estimate the different types of variability, we apply a 20-year LOWESS filter[27] to the data, which should largely remove any short-term variability in the SST that is unrelated to the AMOC. We estimate the interannual variability from the standard deviation of the annual time series minus the 20-year LOWESS-smoothed data. We find the variability in the cold patch to be 0.20 K and 0.19 K from the high-resolution model and observations, respectively. The interannual variability in the warm patch is 0.30 K for both model and observations. We estimate the interdecadal variability from the standard deviation of the 20-year LOWESS-smoothed data minus the linear trend of the smoothed data. The variability is 0.14 K (model) and 0.15 K (observations) for the cold patch, and 0.21 K (model) and 0.18 K (observations) for the warm patch. A discussion of how our results relate to the dominant modes of atmospheric variability in the North Atlantic can be found in Methods.

Conclusions and impacts

We have identified a characteristic SST fingerprint of an AMOC slowdown on the basis of high-resolution model simulations. The fingerprint consists of a cooling in the subpolar gyre region due to reduced heat transport, and a warming in the Gulf Stream region due to a northward shift of the Gulf Stream. This fingerprint is most pronounced during winter and spring, and it is found in the observed long-term

temperature trends, indicating a pronounced weakening of the AMOC since the mid-twentieth century.

We have also defined an improved SST-based AMOC index, which is optimized in its regional and seasonal coverage to reconstruct AMOC changes. Analysis of an ensemble of CMIP5 model simulations confirms that this index can very well reconstruct the long-term trend of the AMOC. We calibrated the observed AMOC decline to be $3 \pm 1 \, \text{Sv}$ (around 15%) since the mid-twentieth century, and reconstructed the evolution of the AMOC for the period 1870–2016. For recent decades, our reconstruction of the AMOC evolution agrees with the results of several earlier studies using different methods, suggesting that our AMOC index can also reproduce interdecadal variations.

Our findings show that in recent years the AMOC appears to have reached a new record low, consistent with the record-low annual SST in the subpolar Atlantic (since observations began in 1880) reported by the National Oceanic and Atmospheric Administration for 2015. Surface temperature proxy data for the subpolar Atlantic suggest that "the AMOC weakness after 1975 is an unprecedented event in the past millennium"[7]. This is consistent with the coral nitrogen-15 data that led Sherwood et al.[28] to conclude that "the persistence of the warm, nutrient-rich regime since the early 1970s is largely unique in the context of the last approximately 1,800 yr". Although long-term natural variations cannot be ruled out entirely[29,30], the AMOC decline since the 1950s is very likely to be largely anthropogenic, given that it is a feature predicted by climate models in response to rising CO_2 levels. This declining trend is superimposed by shorter-term (interdecadal) natural variability.

The AMOC weakening may already have an impact on weather in Europe. Cold weather in the subpolar Atlantic correlates with high summer temperatures over Europe, and the 2015 European heat wave has been linked to the record 'cold blob' in the Atlantic that year[31]. Essentially, low subpolar SSTs were found to favour an air-pressure distribution that channels warm air northwards into Europe. Model simulations further suggest that an AMOC weakening could become the "main cause of future west European summer atmospheric circulation changes"[32], as well as potentially leading to increased storminess in Europe[33]. AMOC weakening has also been connected to above-average sea-level rise at the US east coast[34,35] and increasing drought in the Sahel[19].

Continued global warming is likely to further weaken the AMOC in the long term, via changes to the hydrological cycle, sea-ice loss and accelerated melting of the Greenland Ice Sheet, causing further freshening of the northern Atlantic[36,37]. Given that the AMOC is one of the well documented 'tipping elements' of the climate system, with a defined threshold for collapse[1], it is of considerable concern that the proximity of the Atlantic to this threshold is still poorly known[38–41].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0006-5.

1. Lenton, T. M. et al. Tipping elements in the Earth's climate system. *Proc. Natl Acad. Sci. USA* **105**, 1786–1793 (2008).
2. Rahmstorf, S. Ocean circulation and climate during the past 120,000 years. *Nature* **419**, 207–214 (2002).
3. Masson-Delmotte, V. et al. in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* Ch. 5 (eds Stocker, T. F. et al.) 383–464 (Cambridge Univ. Press, Cambridge, 2013).
4. Smeed, D. A. et al. Observed decline of the Atlantic meridional overturning circulation 2004–2012. *Ocean Sci.* **10**, 29–38 (2014).
5. Dima, M. & Lohmann, G. Evidence for two distinct modes of large-scale ocean circulation changes over the last century. *J. Clim.* **23**, 5–16 (2010).

6. Drijfhout, S., van Oldenborgh, G. J. & Cimatoribus, A. Is a decline of AMOC causing the warming hole above the North Atlantic in observed and modeled warming patterns? *J. Clim.* **25**, 8373–8379 (2012).
7. Rahmstorf, S. et al. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Chang.* **5**, 475–480 (2015); corrigendum **5**, 956 (2015).
8. Booth, B. B. B., Dunstone, N. J., Halloran, P. R., Andrews, T. & Bellouin, N. Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature* **484**, 228–232 (2012). erratum 485, 534 (2012).
9. Saba, V. S. et al. Enhanced warming of the Northwest Atlantic Ocean under climate change. *J. Geophys. Res. Oceans* **121**, 118–132 (2016).
10. Small, R. J. et al. A new synoptic scale resolving global climate simulation using the Community Earth System Model. *J. Adv. Model. Earth Syst.* **6**, 1065–1094 (2014).
11. Delworth, T. L. et al. Simulated climate and climate change in the GFDL CM2.5 high-resolution coupled climate model. *J. Clim.* **25**, 2755–2781 (2012).
12. Olson, R., An, S. I., Fan, Y., Evans, J. P. & Caesar, L. North Atlantic observations sharpen meridional overturning projections. *Clim. Dyn.* https://doi.org/10.1007/s00382-017-3867-7 (2017).
13. Zhang, R. Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation. *Geophys. Res. Lett.* **35**, L20705 (2008).
14. Zhang, R. & Vallis, G. K. The role of bottom vortex stretching on the path of the North Atlantic western boundary current and on the Northern Recirculation Gyre. *J. Phys. Oceanogr.* **37**, 2053–2080 (2007).
15. Zhang, R. et al. Have aerosols caused the observed Atlantic multidecadal variability? *J. Atmos. Sci.* **70**, 1135–1144 (2013).
16. Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, D14 (2003).
17. Stocker, T. F. The seesaw effect. *Science* **282**, 61–62 (1998).
18. Feulner, G., Rahmstorf, S., Levermann, A. & Volkwardt, S. On the origin of the surface air temperature difference between the hemispheres in Earth's present-day climate. *J. Clim.* **26**, 7136–7150 (2013).
19. Defrance, D. et al. Consequences of rapid ice sheet melting on the Sahelian population vulnerability. *Proc. Natl Acad. Sci. USA* **114**, 6533–6538 (2017).
20. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
21. Robson, J., Hodson, D., Hawkins, E. & Sutton, R. Atlantic overturning in decline? *Nat. Geosci.* **7**, 2–3 (2014).
22. Jackson, L. C., Peterson, K. A., Roberts, C. D. & Wood, R. A. Recent slowing of Atlantic overturning circulation as a recovery from earlier strengthening. *Nat. Geosci.* **9**, 518–522 (2016).
23. Frajka-Williams, E. Estimating the Atlantic overturning at 26°N using satellite altimetry and cable measurements. *Geophys. Res. Lett.* **42**, 3458–3464 (2015).
24. Kanzow, T. et al. Seasonal variability of the Atlantic Meridional Overturning Circulation at 26.5°N. *J. Clim.* **23**, 5678–5698 (2010).
25. Latif, M. et al. Is the thermohaline circulation changing? *J. Clim.* **19**, 4631–4637 (2006).
26. Frajka-Williams, E., Beaulieu, C. & Duchez, A. Emerging negative Atlantic multidecadal oscillation index in spite of warm subtropics. *Sci. Rep.* **7**, 11224 (2017).
27. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
28. Sherwood, O. A., Lehmann, M. F., Schubert, C. J., Scott, D. B. & McCarthy, M. D. Nutrient regime shift in the western North Atlantic indicated by compound-specific delta15N of deep-sea gorgonian corals. *Proc. Natl Acad. Sci. USA* **108**, 1011–1015 (2011).
29. Bakker, P., Clark, P. U., Golledge, N. R., Schmittner, A. & Weber, M. E. Centennial-scale Holocene climate variations amplified by Antarctic Ice Sheet discharge. *Nature* **541**, 72–76 (2017).
30. Laepple, T. & Huybers, P. Ocean surface temperature variability: large model–data differences at decadal and longer periods. *Proc. Natl Acad. Sci. USA* **111**, 16682–16687 (2014).
31. Duchez, A. et al. Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 European heat wave. *Environ. Res. Lett.* **11**, 074004 (2016).
32. Haarsma, R. J., Selten, F. M. & Drijfhout, S. S. Decelerating Atlantic meridional overturning circulation main cause of future west European summer atmospheric circulation changes. *Environ. Res. Lett.* **10**, 094007 (2015).
33. Jackson, L. C. et al. Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM. *Clim. Dyn.* **45**, 3299–3316 (2015).
34. Sallenger, A. H., Doran, K. S. & Howd, P. A. Hotspot of accelerated sea-level rise on the Atlantic coast of North America. *Nat. Clim. Change* **2**, 884–888 (2012).
35. Ezer, T. Detecting changes in the transport of the Gulf Stream and the Atlantic overturning circulation from coastal sea level data: the extreme decline in 2009–2010 and estimated variations for 1935–2012. *Global Planet. Change* **129**, 23–36 (2015).
36. Bakker, P. et al. Fate of the Atlantic Meridional Overturning Circulation: strong decline under continued warming and Greenland melting. *Geophys. Res. Lett.* **43**, 12252–12260 (2016).
37. Böning, C. W., Behrens, E., Biastoch, A., Getzlaff, K. & Bamber, J. L. Emerging impact of Greenland meltwater on deepwater formation in the North Atlantic Ocean. *Nat. Geosci.* **9**, 523–527 (2016).
38. Liu, W., Liu, Z. & Brady, E. C. Why is the AMOC monostable in coupled general circulation models? *J. Clim.* **27**, 2427–2443 (2014).
39. Liu, W., Xie, S.-P., Liu, Z. & Zhu, J. Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate. *Sci. Adv.* **3**, e1601666 (2017).
40. Hofmann, M. & Rahmstorf, S. On the stability of the Atlantic meridional overturning circulation. *Proc. Natl Acad. Sci. USA* **106**, 20584–20589 (2009).
41. Buckley, M. W. & Marshall, J. Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: a review. *Rev. Geophys.* **54**, 5–63 (2016).

Author contributions L.C. performed the research and wrote the manuscript together with S.R. S.R. designed the study. A.R. performed the CMIP5 analyses. G.F. helped to interpret the results. V.S. provided the CM2.6 analysis and simulations. All authors discussed the results and provided input to the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data are available for this paper at https://doi.org/10.1038/s41586-018-0006-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.C. or S.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Climate model simulations. The CM2.6 coupled global climate model was developed by the Geophysical Fluid Dynamics Laboratory of the National Oceanic and Atmospheric Administration. It includes an atmospheric general circulation model at an average horizontal resolution of 0.5×0.5 degrees (50 km) and an ocean circulation model at 0.1×0.1 degrees (10 km)[9,11,42]. The ocean has 50 vertical levels and includes a sea-ice model. Two simulations were performed that were both initialized from present-day ocean conditions, followed by a spin-up time of 100 years at constant 1860 CO_2 levels. The control simulation, of 80 years' duration, then maintained CO_2 concentrations at the 1860 level; in the experimental run, by contrast, atmospheric CO_2 increased by 1% per year over 70 years until it doubled, and then remained at this level for another 10 years. Given the extremely high computational cost of this model (approximately one day per one year of simulation on a high-performance computer), no further simulations are available.

Definition of the AMOC index. We define the AMOC index I_{AMOC} as the difference between the mean SST of the geographic region that is most sensitive to a reduction in the AMOC (the subpolar gyre region, sg) and that of the whole globe:

$$I_{AMOC} = \overline{SST_{sg}} - \overline{SST_{global}}$$

Rather than including the whole year, we instead use only the winter and spring months (November to May), because the AMOC signal found in the SST is most pronounced during these seasons (see Fig. 4). Thus the AMOC index for a certain year is defined as the mean SST in the subpolar gyre region for the following November−May season, minus the global mean SST for that season.

Definition of the subpolar gyre region. To define the region used to calculate the AMOC index (shown in the inset of Fig. 3), we assumed that SST differences in the subpolar North Atlantic relative to the global mean SST are dominated by variations in the AMOC. For this study, we determined this region by combining normalized linear SST trends from both the HadISST dataset and the high-resolution CM2.6 model run, as shown in Fig. 2. Grid cells that show relative cooling in either the observations or the model were included in the definition. The region is large (compared, for example, with that used in ref. [7]), which has the advantage that it should cover most of the area in which the heat transported northwards by the AMOC is vented to the atmosphere in the observations and in the models, especially considering that the exact location of heat release is, to some degree, model-dependent. The exact coordinates of the region are available in a public data repository (see Data availability).

Definition of the Gulf Stream region. Similar to the subpolar gyre region, the Gulf Stream region is defined as the region that covers the above-average long-term warming east of the US coast that results from an AMOC slowdown in both observations and model (see inset of Fig. 3). Thus, the terms Gulf Stream region and subpolar gyre region do not refer directly to ocean circulation features, but rather to SST features. The exact coordinates of the region are available in a public data repository (see Data availability).

AMOC effects on Gulf Stream separation point and DWBC strength. We link the extreme warming observed along the US coast to the Gulf Stream shifting northwards and closer to shore as a consequence of an AMOC slowdown. For the MOM4 ocean model, it has been shown that the correct separation point of the Gulf Stream is achieved through a reasonable representation of the DWBC[14]. Furthermore it has been shown that, for this model, a weakening of the AMOC is accompanied by a weakening of the DWBC and that both are followed by a northward shift of the mean Gulf Stream path[43,44]. The combination of these results indicates that, in the model run, the observed warming is indeed due to a weakened AMOC that leads to a weakened DWBC, a weakened northern recirculation gyre and a northern shift of the Gulf Stream separation point. To test this, we compared the evolution of the Gulf Stream path (represented by the Gulf Stream index—that is, the mean latitude of the 15 °C isotherm at a 200-m depth in the Northwest Atlantic, between 75° W and 55° W[44]) with the AMOC strength at 26° N in the CM2.6 control run and the CO_2-doubling run (Extended Data Fig. 4a).

We compared the AMOC strength to the summed southward deep-ocean transport (between depths of 1,000 m and 4,000 m) at 40° N in the region between the coast and 65° W, for the CM2.6 control run and the CO_2-doubling run (Extended Data Fig. 4b). We found that the DWBC in the model indeed weakens as the AMOC slows down, and by a very similar amount (around 3.5 Sv). We calculated the DWBC at this latitude because it is just north of the region where the Gulf Steam and DBWC cross in the control run, and is thus the area where the northern recirculation gyre forms, which forces the Gulf Stream to deflect from the coast. These analyses confirm that the AMOC weakening in the model is indeed accompanied by a weakened DWBC and a northerly shift of the Gulf Stream path.

Analysis of additional observational datasets. For this study, we analysed seven available SST data products. All of them show the fingerprint of the AMOC,

namely, the cold patch in the subpolar Atlantic and, to a greater or lesser extent, the enhanced warming inshore of the Gulf Stream (Extended Data Fig. 5). Details of the different datasets are given in Extended Data Table 1. Different choices of processing steps lead to distinctions in the representation of spatial and temporal variability in the datasets. We focused on the dataset with the best spatial resolution and advanced quality control, the HadISST data. Although the ERSST data are also quality-controlled, the use of empirical orthogonal teleconnections for post-processing leads to a smoothing of the SST signal in the spatial domain (unwanted for this study). The bias adjustments and quality-control procedures used for the likewise high-resolution COBE dataset are not as advanced as those used for the HadISST and ERSST data. The SODA data are an ocean reanalysis product, that is, they are based on model simulations with data assimilation.

Significance of the 1870–2016 trends. To illustrate the significance of the 1870–2016 linear trends, we compare the distribution of the long-term trends for all grid cells between 60° S and 75° N with the distribution of trends for the grid cells in the subpolar gyre region and with the grid cells in the Gulf Stream region (defined in the inset of Fig. 3). (We exclude the sea-ice-covered regions because they are expected to show no temperature trend, consistent with the assumption that SSTs remain close to freezing point there.) Extended Data Fig. 3 shows the global distributions of relative SST trends for the HadISST data and the CO_2-doubling run of the CM2.6 model. Assuming a constant bin size of 0.2, we determined the 5% and 95% quantiles. The medians of the subpolar gyre and Gulf Stream regions lay in all cases within the lowest and highest 5% of the trends. The median of the Gulf Stream region in the HadISST data is 2.4 (that is, the warming here is 2.4 times larger than the global SST warming), higher than 96% of the SST trends; the median of the subpolar gyre region is −0.17, and thus among the lowest 3% of the trends. In the CO_2-doubling run of the CM2.6 model, the AMOC fingerprint regions are even greater outliers, presumably because the larger global-warming signal and associated greater AMOC weakening result in a better signal-to-noise ratio. The median of the Gulf Stream region in the models is 2.4, higher than 98% of the SST trends, and the median of the subpolar gyre region is −0.25, among the lowest 1% of the trends.

Relation between the AMOC index and the overturning strength. To assess and calibrate the relation between changes in the AMOC index and the AMOC strength, we examined the AMOC index and AMOC simulations performed using 15 models in the context of CMIP5 for the historical (1870–2005) climate, extended to 2016 using simulations of the RCP8.5 scenario. To assess whether the models have a reasonable representation of the AMOC, we compared the mean maximum AMOC at 26° N for the model years 2005–2014 with the mean of the observed AMOC at around 26° N during that period (16.8 Sv; see Extended Data Table 1). We chose models with mean maximum AMOCs of 16.8 ± 10.0 Sv; this excluded the NorESM1-M and Nor-ESM1-ME models. We further excluded the GISS-E2-R model because it is an outlier with a very unrealistic deep mixed layer that reaches down to the sea floor in most of the subpolar Atlantic[45].

Total-least-squares fit. To test the relation between our AMOC index and the AMOC strength, we performed a total-least-squares fit (also known as an orthogonal regression, because the error in both variables is minimized—that is, the error is orthogonal to the regression line). The full regression equation is:

$$Y = 3.8\ \text{Sv K}^{-1} \times X + 0.1\ \text{Sv per century}$$

where X is the trend in AMOC indices, in kelvins per century, and Y is the corresponding trend in AMOC strength, in sverdrups per century.

Sensitivity to extension of the subpolar gyre region. The region chosen as the subpolar gyre region is, on average, largely free of sea ice (to analyse this, we compared the region with the average November–May sea-ice cover from the HadISST data). To explore how partly ice-covered areas influence the index, we limited the region to ice-free areas (determined by the maximum sea-ice cover for the November–May season from 1870 to 2016), and compared the resulting index with our original AMOC index (Extended Data Fig. 7). This shows some differences in the year-to-year variations, but the longer-term trend, especially in the last decades, is hardly affected at all. Thus we conclude that sea ice does not affect our AMOC index.

Comparison with a previous AMOC index. Rahmstorf et al.[7] used a different region and different data (annual HadCRUT4 SSTs minus their annual Northern Hemispheric mean, both land and ocean) to obtain the AMOC index. We calculate the AMOC index relative to the global mean SST; however, as our comparison of the two indices shows, the index is not sensitive to this choice (Extended Data Fig. 6). Rahmstorf et al.[7] also determined the conversion factor between their AMOC index and the actual AMOC by using only one model, MPI-ES-MR. We updated their AMOC index with the latest data and compared it with the AMOC slowdown determined herein (Extended Data Fig. 6). The results that we obtained with both index definitions are highly consistent on the multidecadal timescale of interest.

Link to empirical modes of variability. Two main modes of variability have been defined in the North Atlantic, primarily on the basis of empirical data: the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO). The former describes atmospheric variability, with an index based on the surface pressure gradient[46], whereas the latter describes SST variability relative to the global mean—similar to our AMOC index, but including Atlantic SSTs down to the Equator. Both NAO and AMO indices show a correlation with our AMOC index (Extended Data Figs. 8, 9).

For the AMO index this is not surprising, given that it has the subpolar SST data in common with our AMOC index. However, the usefulness of the AMO index is limited by the fact that it conflates subpolar SST variability and tropical SST variability into one index[47]. For our purpose of using SSTs to deduce AMOC variations, this degrades the signal-to-noise ratio. Furthermore, it can be seen that the decadal variations in our AMOC index are similar to those of the AMO index (Extended Data Fig. 8b), which is in accordance with other studies showing that the time evolution of the AMO can at least partly be explained by changes in Atlantic Ocean currents[48,49]. Yet because the AMO conflates two regions with different long-term trends—that is, the subpolar North Atlantic, which is cooling, and the tropics and subtropics, which are warmer with temperature trends at or above the rate of the global mean (Fig. 2)—it does not show the 1870–2016 negative trend that is clearly visible in our AMOC index (Extended Data Fig. 8a).

The NAO index is more useful, as it can be used to study the relationship between atmospheric-pressure variability and North Atlantic SSTs. We find a clear negative correlation with $R = -0.54$ between the decadally smoothed time series of the AMOC and the NAO indices, which occurs when the AMOC leads the NAO by three years (see Extended Data Fig. 9b). This negative correlation, and the fact that a pronounced cooling in the subpolar North Atlantic has been shown to be followed by a positive phase of the NAO[50], suggests that on interdecadal timescales the AMOC at least partially drives NAO changes via changes in North Atlantic SSTs, rather than the other way round. Consistent with this, the NAO index shows a positive trend for 1870–2016 (Extended Data Fig. 9a). A positive NAO, on the other hand, helps to extract heat from the subpolar ocean through enhanced westerly winds over that region, cooling SSTs, enhancing convection and increasing ocean density[51]. This acts as a negative feedback on an AMOC weakening. Such a delayed negative feedback could either dampen the AMOC response or lead to oscillatory behaviour. Further investigation of this linkage is beyond the scope of this study. Nevertheless, our work supports the importance of ocean circulation to variations in the North Atlantic SST pattern, which has been highlighted previously[52,53].

Code availability. Code for running the CM2.6 experiment is available from http://www.gfdl.noaa.gov/. Scripts for analysing the data are available from the corresponding authors upon reasonable request.

Data availability. The SST datasets analysed here are publicly available; detailed information is given in Extended Data Table 1. The CMIP5 model output is available from https://esgf-node.llnl.gov/projects/cmip5/. The CM2.6 model output is available from V.S. (vincent.saba@noaa.gov) upon reasonable request. The exact definitions of the subpolar gyre and Gulf Stream region, as well as the SST anomalies of these regions, are available in a public data repository: http://www.pik-potsdam.de/~caesar/AMOC_slowdown/. The data for the GloSea reanalysis were provided by L. Jackson[22]. The data for the reconstruction from satellite altimetry and cable measurements were provided by E. Frajka-Williams[23].

RAPID data are available from http://www.rapid.ac.uk/rapidmoc/rapid_data/datadl.php.

42. Griffies, S. M. et al. Impacts on ocean heat from transient mesoscale eddies in a hierarchy of climate models. *J. Clim.* **28**, 952–977 (2015).
43. Zhang, R. & Vallis, G. K. Impact of great salinity anomalies on the low-frequency variability of the North Atlantic climate. *J. Clim.* **19**, 470–482 (2006).
44. Sanchez-Franks, A. & Zhang, R. Impact of the Atlantic meridional overturning circulation on the decadal variability of the Gulf Stream path and regional chlorophyll and nutrient concentrations. *Geophys. Res. Lett.* **42**, 9889–9897 (2015).
45. Heuzé, C. North Atlantic deep water formation and AMOC in CMIP5 models. *Ocean Sci.* **13**, 609–622 (2017).
46. Visbeck, M. H., Hurrell, J. W., Polvani, L. & Cullen, H. M. The North Atlantic Oscillation: past, present, and future. *Proc. Natl Acad. Sci. USA* **98**, 12876–12877 (2001).
47. Frankignoul, C., Gastineau, G. & Kwon, Y.-O. Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation. *J. Clim.* **30**, 9871–9895 (2017).
48. Zhang, R., Delworth, T. L. & Held, I. M. Can the Atlantic Ocean drive the observed multidecadal variability in Northern Hemisphere mean temperature? *Geophys. Res. Lett.* **34**, L02709 (2007).
49. O'Reilly, C. H., Huber, M., Woollings, T. & Zanna, L. The signature of low-frequency oceanic forcing in the Atlantic Multidecadal Oscillation. *Geophys. Res. Lett.* **43**, 2810–2818 (2016).
50. Gastineau, G. & Frankignoul, C. Influence of the North Atlantic SST variability on the atmospheric circulation during the twentieth century. *J. Clim.* **28**, 1396–1416 (2015).
51. Delworth, T. L. & Zeng, F. The impact of the North Atlantic Oscillation on climate through its influence on the Atlantic Meridional Overturning Circulation. *J. Clim.* **29**, 941–962 (2016).
52. Delworth, T. L. et al. The central role of ocean dynamics in connecting the North Atlantic Oscillation to the extratropical component of the Atlantic Multidecadal Oscillation. *J. Clim.* **30**, 3789–3805 (2017).
53. Zhang, R. On the persistence and coherence of subpolar sea surface temperature and salinity anomalies associated with the Atlantic multidecadal variability. *Geophys. Res. Lett.* **44**, 7865–7875 (2017).
54. Huang, B. et al. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
55. Huang, B. et al. Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: upgrades and Intercomparisons. *J. Clim.* **28**, 911–930 (2015).
56. Smith, T. M., Reynolds, R. W., Peterson, T. C. & Lawrimore, J. Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Clim.* **21**, 2283–2296 (2008).
57. Kaplan, A., Cane, M. A., Kushnir, Y., Clement, A. C., Blumenthal, M. B. & Rajagopalan, R. Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res. Oceans* **103**, 18567–18589 (1998).
58. Carton, J. A. & Giese, B. S. A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA). *Mon. Weath. Rev* **136**, 2999–3017 (2008).
59. Hirahara, S., Ishii, M. & Fukuda, Y. Centennial-scale sea surface temperature analysis and its uncertainty. *J. Clim.* **27**, 57–75 (2014).
60. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, L12704 (2006).
61. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
62. Intergovermental Panel on Climate Change. Climate Change 2013: The Physical Science Basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, 2013).
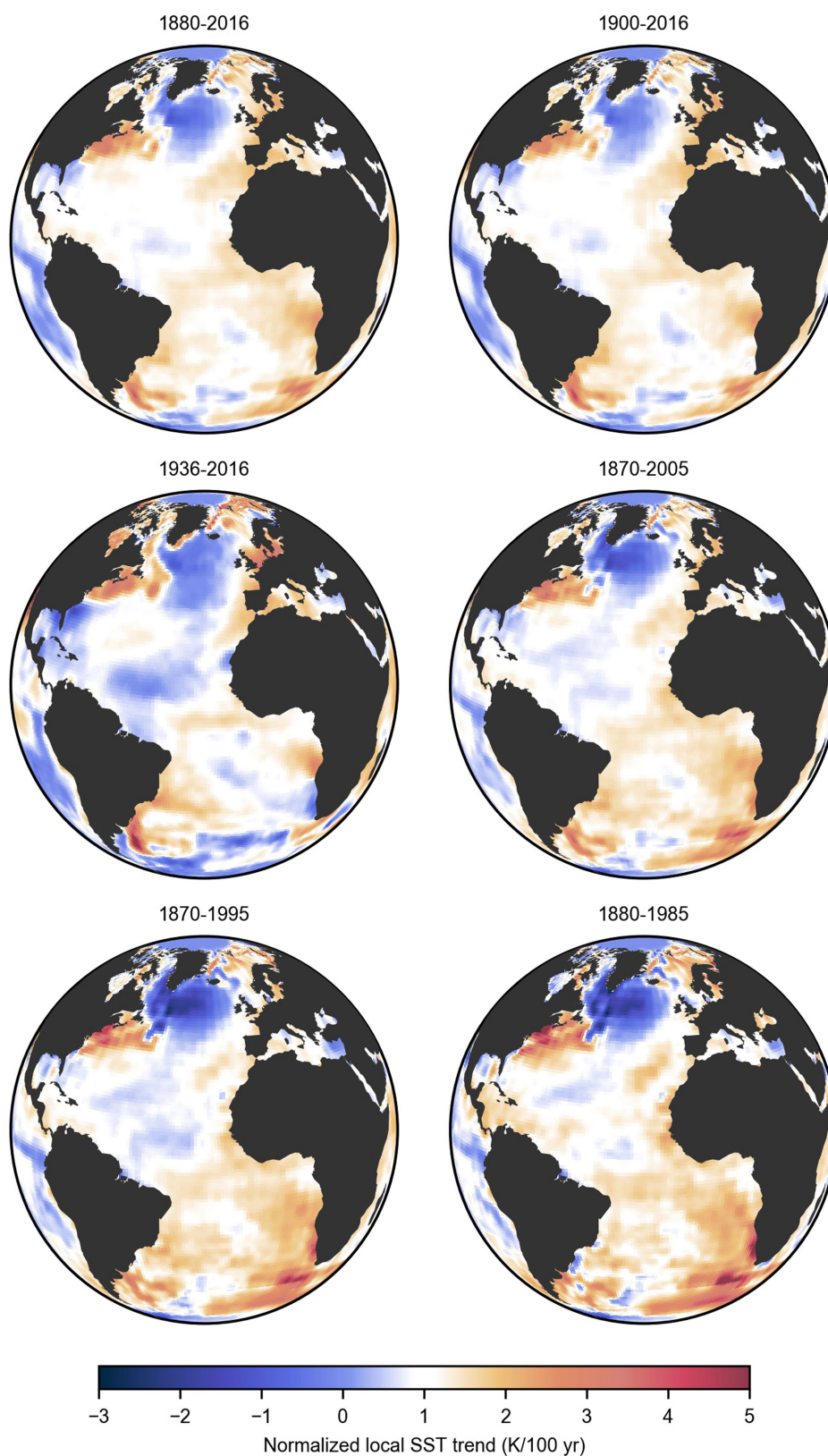
Extended Data Fig. 1 | Normalized SST trends in the HadISST data for different time periods. Observed linear SST trends (using annual HadISST data), calculated for different timespans to test the robustness of the linear SST trend pattern to the starting and ending years of the timespan. The pattern is normalized with the respective global mean SST trend. Regions that show below-average warming or cooling are in blue; regions that show above-average warming are in red.
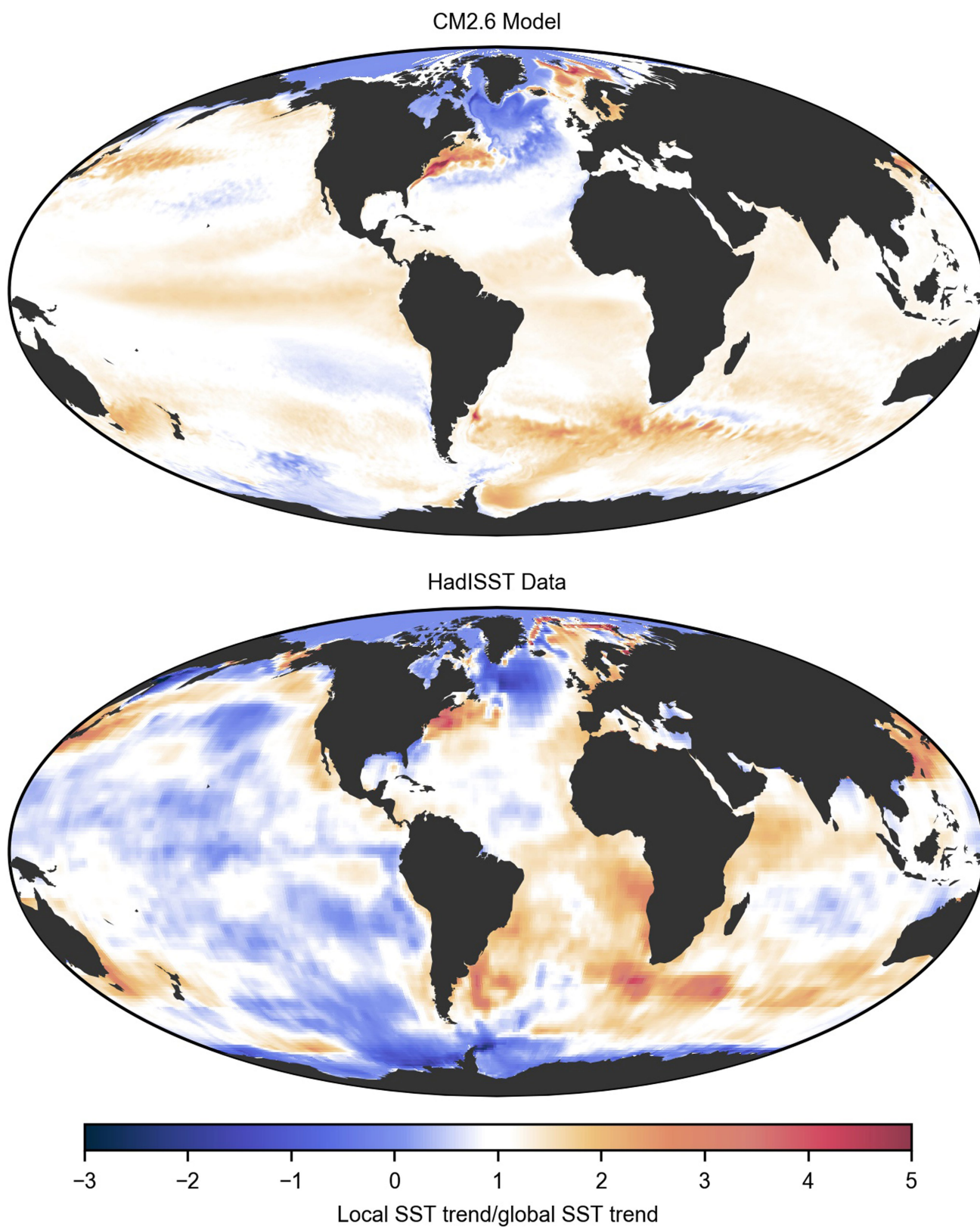
Extended Data Fig. 2 | Comparison of global normalized SST trends. Linear SST trends during a CO_2-doubling experiment using the GFDL CM2.6 climate model (top), and observed trends during 1870–2016 (HadISST data, bottom), both normalized with the respective global mean SST trends and using data from the November–May season. Regions that show cooling or below-average warming are in blue; regions that show above-average warming are in red. Note again that owing to the much greater climate change in the CO_2-doubling experiment, the signal-to-noise ratio for the modelled SST trends is better than that for the observations, and thus the noise level is suppressed by the normalization.
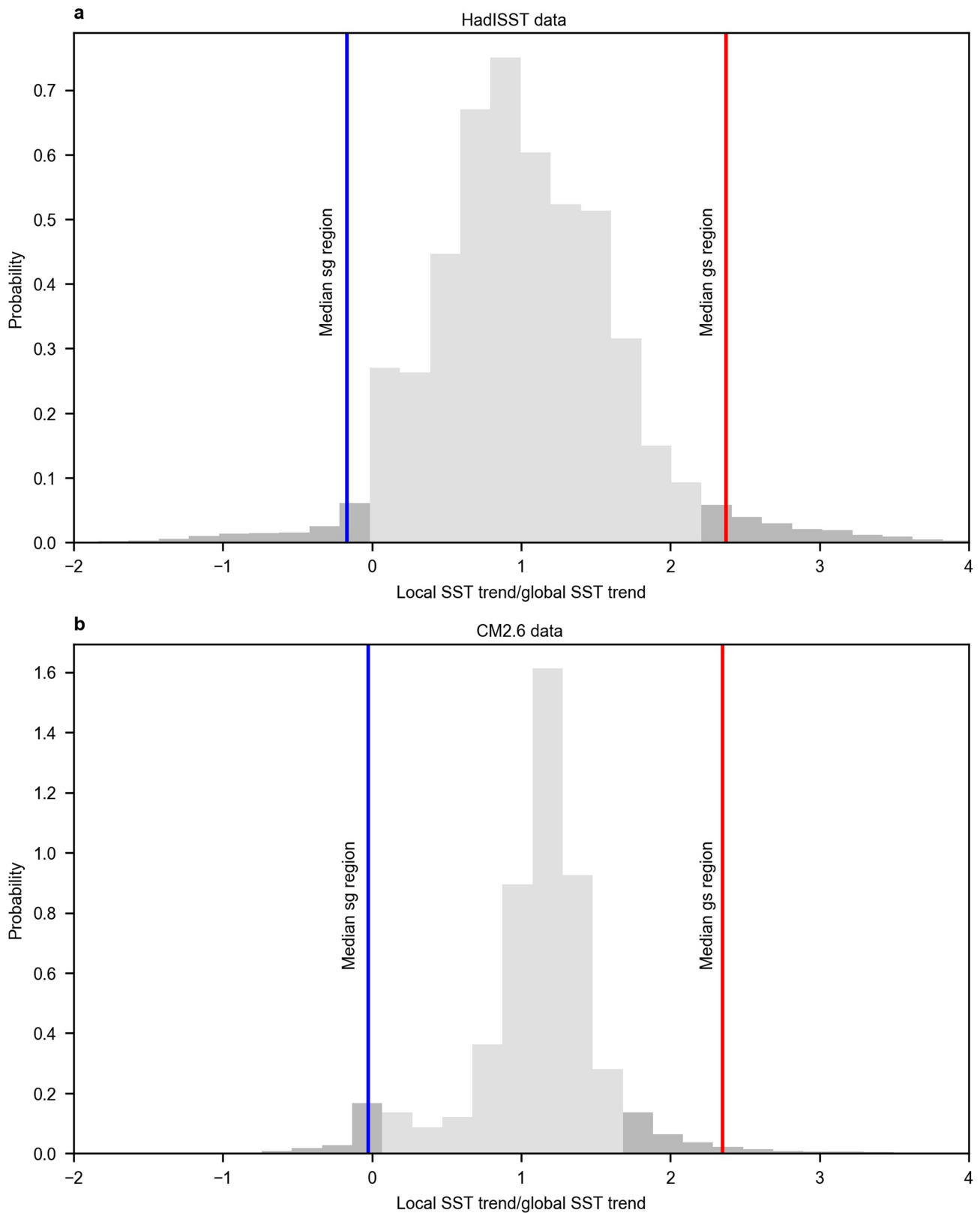
Extended Data Fig. 3 | Histograms showing the distribution of the normalized longer-term trends. a, The distribution (grey bars) of all local trends, normalized to the global trends, from the HadISST data for 1870–2016, for latitudes between 60° S and 75° N. The distribution is located around $\mu = 1$ with a standard deviation of $\sigma = 0.66$ (grey bars). The 5th and 95th percentiles are marked in darker grey. The distribution of the 1870–2016 trends for grid cells assigned to the subpolar gyre regions is shifted to lower or even negative values, with a median of $\tilde{x}_{sg} = -0.17$ (blue). The distribution of trends for grid cells in the Gulf Stream region are shifted to

higher values, with a median of $\tilde{x}_{gs} = 2.4$ (red). The distributions are normalized to account for the different sample sizes of global, subpolar gyre and Gulf Stream regions. **b**, As for panel **a**, but for the CO_2-doubling run of the CM2.6 model, with $\mu = 1.1$, $\sigma = 0.48$, $\tilde{x}_{sg} = -0.02$ and $\tilde{x}_{gs} = 2.4$. The standard deviations of the model data are expected to be smaller than those of the observations because of the larger climate-change signal by which the model data are normalized; this reduces the 'noise' of short-term variability relative to the climate signal.
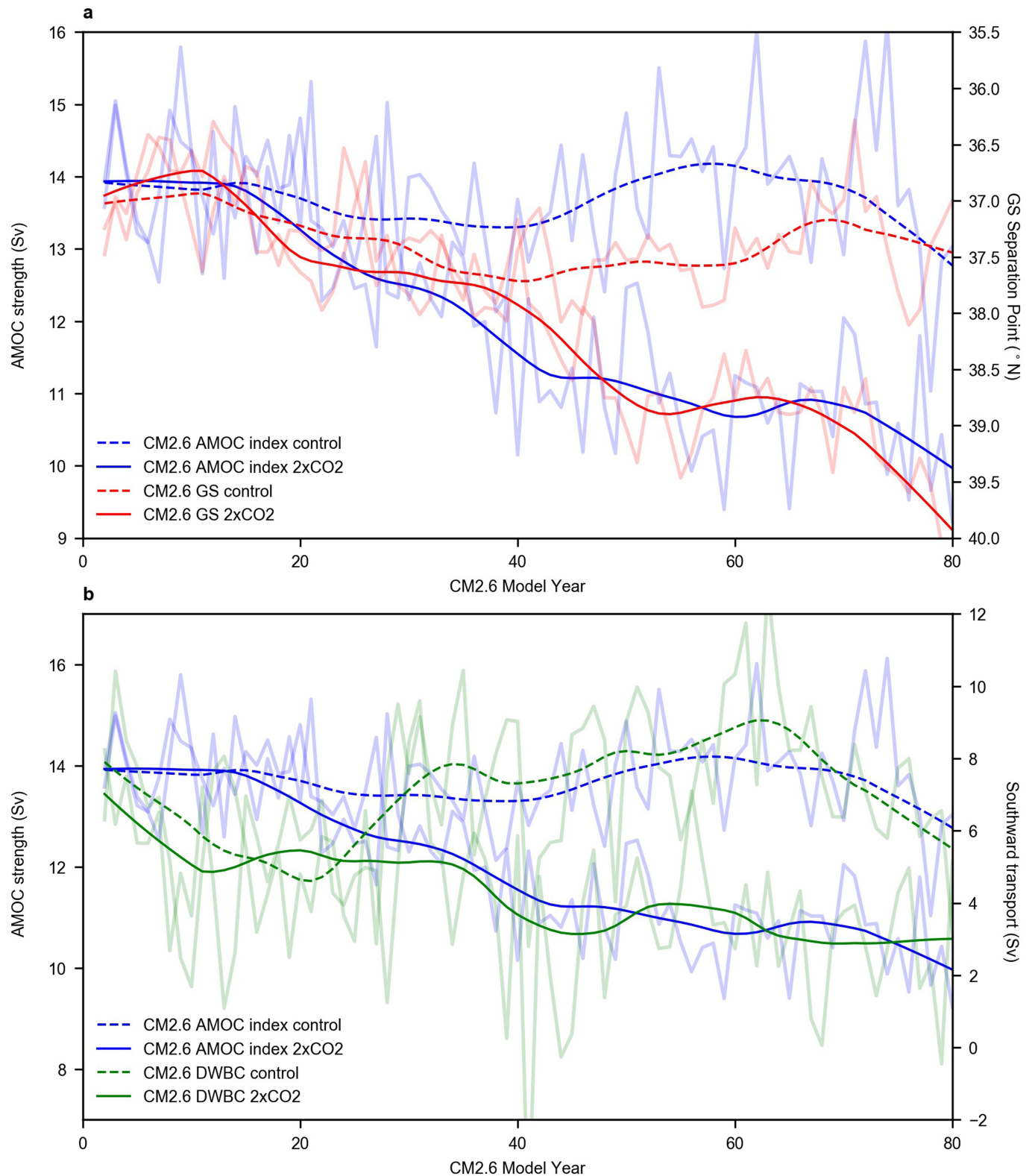
Extended Data Fig. 4 | Influence of the AMOC on the separation point of the Gulf Stream. a, The evolution of the Gulf Stream (GS) separation point compared with the AMOC strength in the CM2.6 control and CO$_2$-doubling runs, as indicated by the Gulf Stream index[44]. The graph shows a link between a weaker AMOC and a northward shift of the separation point. **b**, Time series of the southward transport of the deep ocean current (summed between depths of 1,000 m and 4,000 m) at 40° N in the region between the US coast and 65° W (see Methods), showing a weakening DWBC during the CO$_2$-doubling experiment. The thin lines show annual values, the thick lines show the 20-year LOWESS-smoothed values.
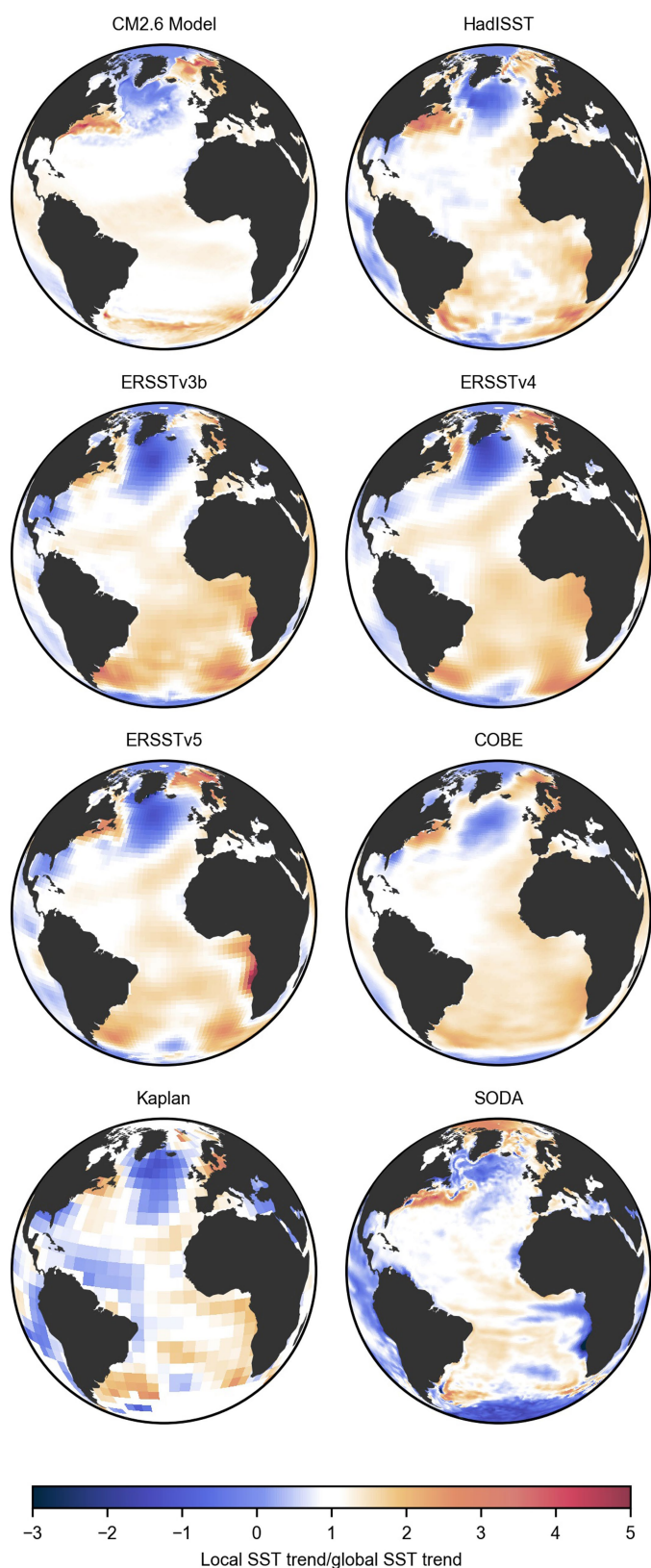
Extended Data Fig. 5 | Linear SST trends from a CO_2-doubling experiment using the GFDL CM2.6 climate model, and observed long-term trends from different SST data products, normalized with the respective global mean SST trends. The trend from 1870 to 2016 was calculated using those datasets that provide data until the present (HadISST[16], ERSSTv5[54], ERSSTv4[55], ERSSTv3b[56] and Kaplan[57]). Otherwise, it was calculated from 1870 to the end of the available time period (SODA[58] and COBE[59]; see Extended Data Table 1). The SODA data are given for a depth of 5 m instead of the surface; thus, the long-term trend differs for regions with ice cover. For the SODA data, the normalization was adjusted with surface SST data instead of the data at a 5-m depth, to make this dataset comparable to the others. All datasets show a prominent cooling in the subpolar gyre region; the high-resolution data (HadISST, COBE and SODA) also show pronounced warming in the Gulf Stream region.
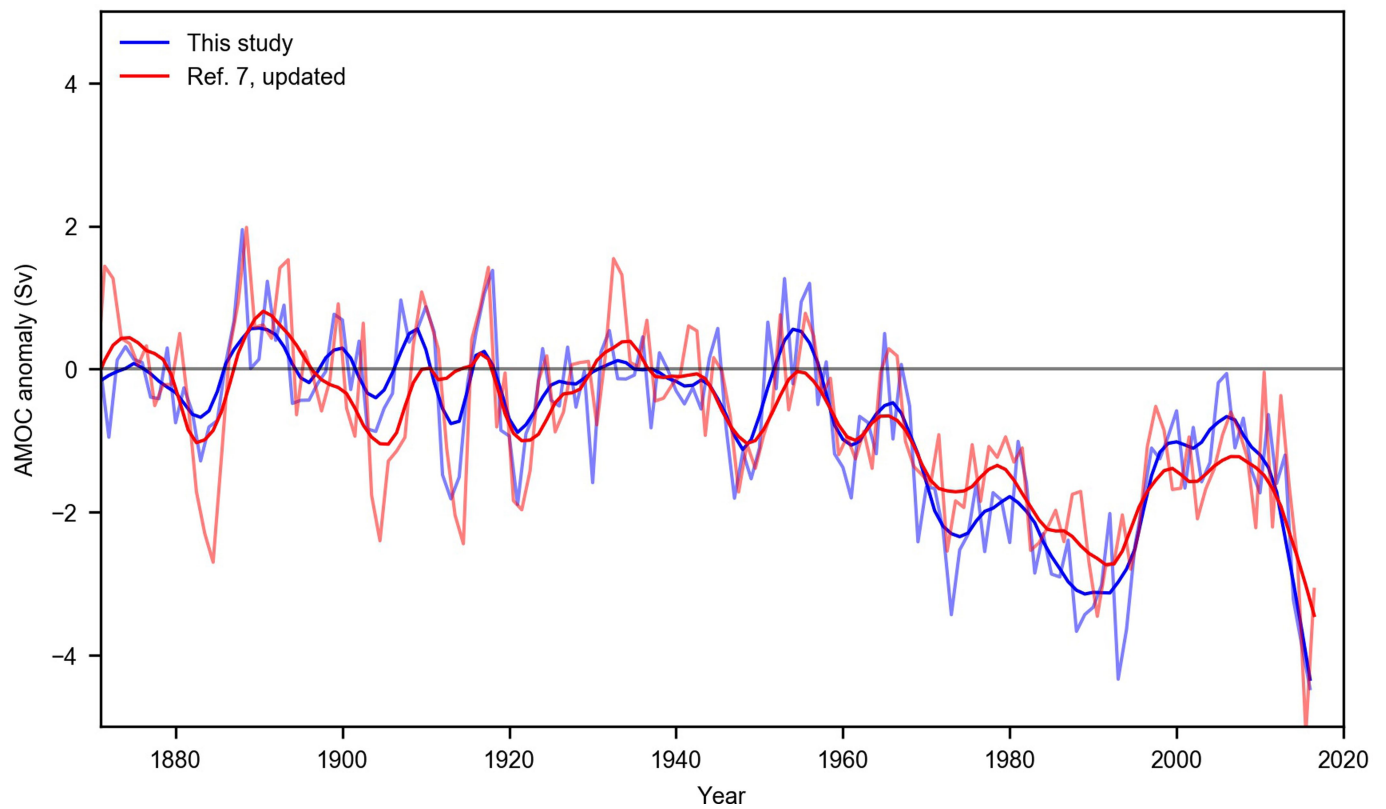
Extended Data Fig. 6 | Time series of the AMOC anomaly for two definitions of the AMOC index. We calculated the AMOC anomaly from two AMOC indices and two model-based conversion factors. In red is the AMOC anomaly as defined by Rahmstorf et al.[7] (HadCRUT4 data), updated with the latest data to 2016. In blue is the AMOC anomaly as defined herein (HadISST data). Thick lines are smoothed by a 10-year LOWESS filter. This smoothing filter is lower than that used in Fig. 6, in order to compare and show the two indices with a higher time resolution.
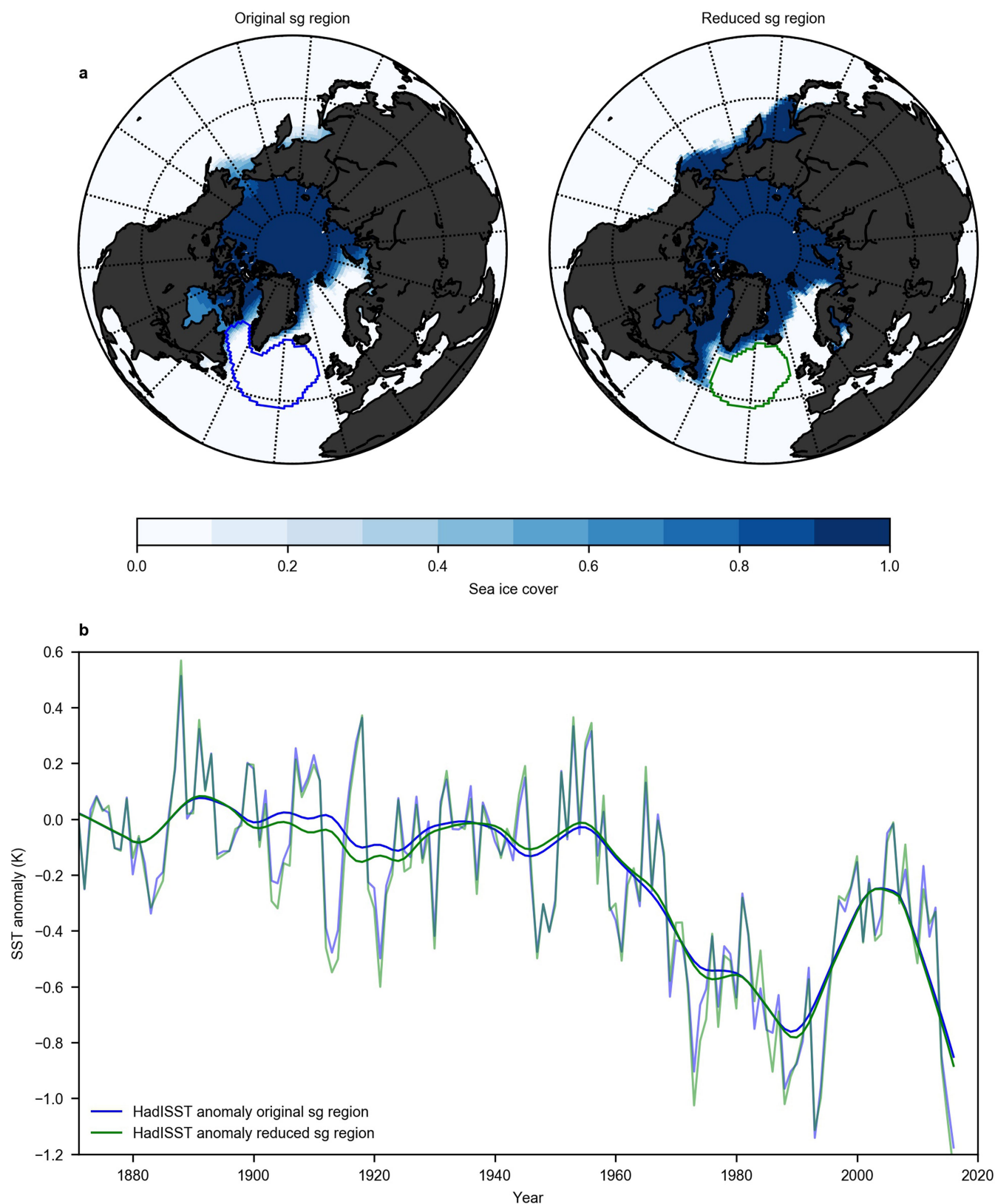
Extended Data Fig. 7 | Sensitivity to the extension of the subpolar gyre region regarding sea-ice cover. a, Left panel, our original subpolar gyre region (blue outline) and the average November–May sea-ice cover from 1870 to 2016 (blue shading, from HadISST data). Right panel, a reduced subpolar gyre region (green outline) that is always ice-free, compared with the maximum sea-ice cover for the November–May season from 1870 to 2016. **b**, Comparison of the AMOC indices based on these two regions. The thin lines show annual values, the thick lines show the 20-year LOWESS-smoothed values.
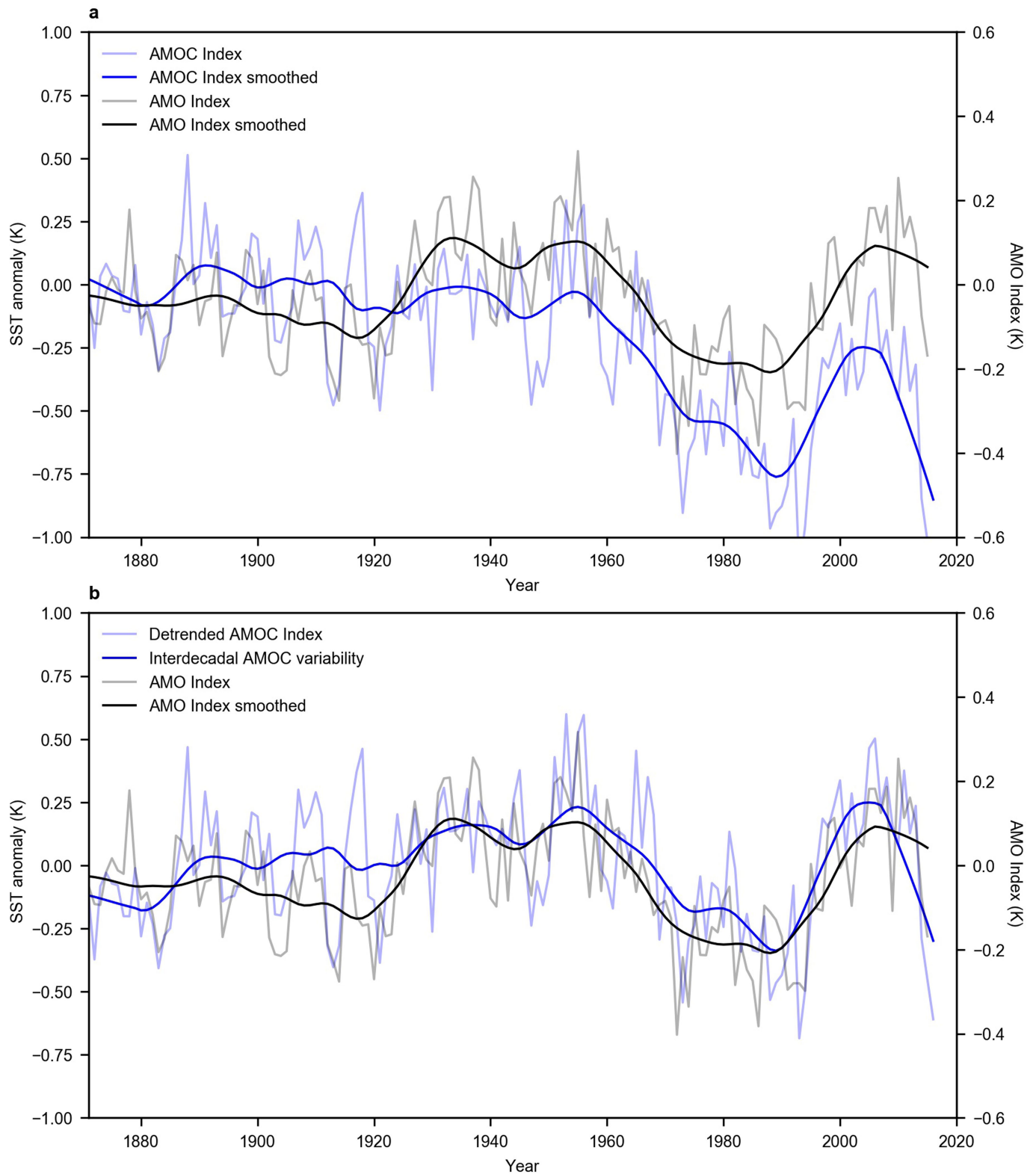
Extended Data Fig. 8 | Comparison of interdecadal variability of the AMOC index and the AMO index. a, We calculated the AMO index from the HadISST dataset after Trenberth and Shea[60]. This index is defined as the weighted mean SST over the North Atlantic (0° N to 80° N), relative to the mean SST from the period 1901–1970, but with the global mean SST (averaged over the global oceans from 60° S to 60° N) removed. The thin lines show annual values, the thick lines indicate the 20-year LOWESS-smoothed values. We show our AMOC index for comparison. **b**, As for panel **a**, but here the AMO index is compared with the interdecadal variability of our AMOC index—that is, the detrended 20-year LOWESS-smoothed index. The comparison shows that the AMO index has similar interdecadal variability to the AMOC index but is lacking the climatic trend found in the latter.
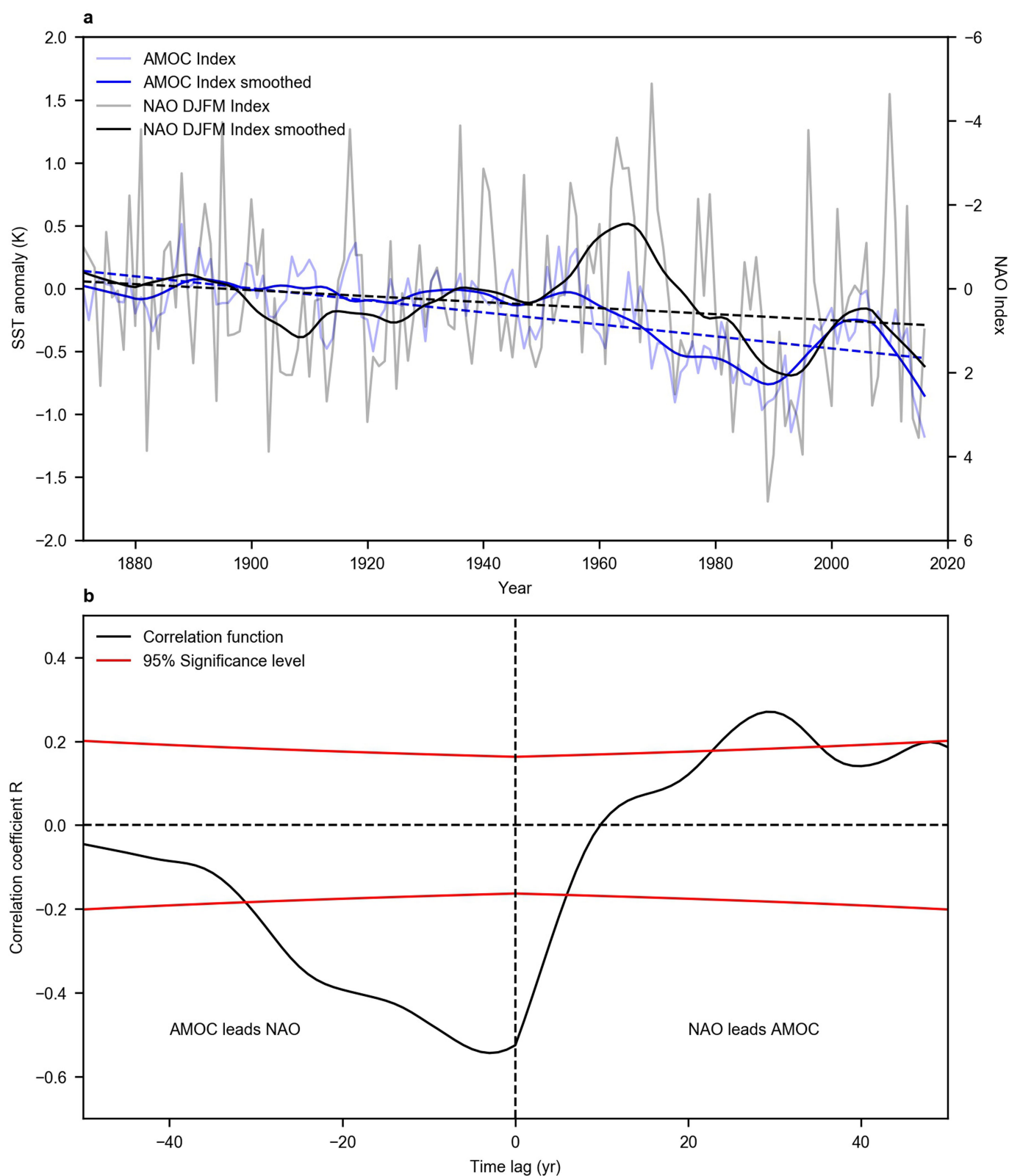
Extended Data Fig. 9 | Comparison of AMOC and NAO. a, Comparison of our AMOC index with the interdecadal variability in the NAO index (after Hurrell[61]), calculated as the sea-level pressure at the Lisbon station minus the sea-level pressure at the Stykkisholmur/Reykjavik station for the months December to March (DJFM). The thin lines show annual values, the thick lines show the 20-year LOWESS-smoothed values. The linear trend over the whole time period is shown with dashed lines. **b**, Lagged cross-correlation between the AMOC index and the NAO index shows that peak negative correlation occurs when the AMOC leads the NAO by three years, with $R = -0.54$. The red lines mark the 95% significance level.

Extended Data Table 1 | Detailed data and model information

Data set	HadISST1[16]	ERSSTv4[55]	ERSSTv3b[56]	SODA[58]	COBE[59]	Kaplan[57]	ERSSTv5[54]
Full Name	Merged Hadley-NOAA/Oi SST	NOAA Extended Reconstruction SST version 4	NOAA Extended Reconstruction SST version 3b	Simple Ocean Data Assimilation (2.2.4)	Centennial in situ Observation-Based Estimates	Kaplan Extended SST v2	NOAA Extended Reconstruction SST version 5
Resolution	1.0°latx1.0°lon, monthly	2.0°latx2.0°lon, monthly	2.0°latx2.0°lon, monthly	0.5°latx0.5°lon, 40 vertical levels, monthly	1.0°latx1.0°lon, monthly	5.0°latx5.0°lon, monthly	2.0°latx2.0°lon, monthly
Period of Record	01/1870 – present	01/1854 – present	01/1854 – present	01/1870 – 12/2010	01/1850 – 12/2015	01/1856 – present	01/1854 – present
Input Data	SST data are taken from the Met Office Marine Data Bank (MDB), ICOADS and Global Telecommunication System (GTS) data.	ICOADS Release 2.5 and NCEP GTS data.	ICOADS Release 2.4 and NCEP GTS data.	Assimilation of a model forecast and observation data (including surface temperature and salinity observations of various types, and nighttime infrared satellite SST data).	ICOADS Release 2.5.	MOHSST5 version of the GOSTA data set from the UK MET office.	ICOADS Release 3.0 and NCEP GTS data
Processing steps	Two stage reduced-space optimal interpolation (RSOI) procedure, followed by superposition of quality-improved gridded observations onto the reconstructions.	Updated quality control and bias correction with Marine Nighttime Air temperatures. Updated Fitting with Empirical Orthogonal Teleconnections (EOTs).	Quality control and bias correction with Marine Nighttime Air temperatures. Fitting with Empirical Orthogonal Teleconnections (EOTs).	Assimilation product with quality control.	Gridding via optimal interpolation (OI) as well as bias-adjustment and quality control.	Adjustments and Filling via EOF projection, Optimal Interpolation (OI), Kalman Filter (KF) forecast, KF analysis, and an Optimal Smoother (OS).	Improved SST spatial and temporal variability by reducing spatial filtering in training the reconstruction EOTs, removing high-latitude damping in EOTs, and adding 10 more EOTs in the Arctic. Switch from using Nighttime Marine Air Temperature (NMAT) as a reference to buoy-SST as a reference in correcting ship SST biases.
Source	http://www.metoffice.gov.uk/hadobs/hadisst/data/download.html	https://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v4	https://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v3b	http://dsrs.atmos.umd.edu/DATA/soda_2.2.4/	https://www.esrl.noaa.gov/psd/data/gridded/data.cobe2.html	https://www.esrl.noaa.gov/psd/data/gridded/data.kaplan_sst.html	https://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v5/netcdf/
Download date	10/21/2016 (update on 04/04/2017)	11/19/2016 (update 06/07/2017)	01/23/2017	01/23/2017 (updated 02/21/2017)	02/01/2017	02/01/2017	08/14/2017

Model	Institute	Average resolution of oceanic model	Mean AMOC 26 °N [Sv]	AMOC index decline [K/century]	AMOC decline [Sv/century]
CanESM2	CCCMA (Canada)	256 × 192	14.76	-0.16	-0.1
CCSM4	NCAR (USA)	Nominal 1° (1.125° in longitude, 0.27–0.64° variable in latitude)	17.72	-0.57	-2.2
CESM1-BGC	NCAR (USA)	Nominal 1° (1.125° in longitude, 0.27–0.64° variable in latitude)	18.63	-0.71	-2.7
CESM1-CAM5-1-FV2	NCAR (USA)	same as CESM1-BGC	19.02	-0.19	-0.5
CESM1-CAM5	NCAR (USA)	same as CESM1-BGC	18.68	-0.21	-0.5
CNRM-CM5	CNRM-CERFACS (France)	0.7° on average ORCA1	12.33	-0.54	-2.1
GFDL-ESM2M	NOAA GFDL (USA)	1° tripolar 360 × 200L50	18.09	-0.78	-2.6
GISS-E2-R	NASA GISS (USA)	0.2 to 1° latitude × 1° longitude	18.22	0.02	-2.1
INMCM4	INM (Russia)	1 × 0.5° in longitude and latitude generalized spherical coordinates with poles displaced	17.73	-0.40	-2.5
MPI-ESM-LR	MPI-M (Germany)	average 1.5° GR15	18.42	-0.54	-2.0
MPI-ESM-MR	MPI-M (Germany)	approx. 0.4° TP04	16.67	-0.06	-0.6
MRI-CGCM3	MRI (Japan)	1.0° × 0.5°	14.42	-0.08	-0.1
MRI-ESM1	MRI (Japan)	1.0° × 0.5°	14.90	-0.10	-0.4
NorESM1-ME	NCC (Norway)	1.125° along the equator	29.56	-0.45	-0.1
Nor-ESM1-M	NCC (Norway)	1.125° along the equator	28.69	-0.27	-0.7

Overview of the spatial and temporal resolution, period of record, input data, processing steps and sources of the 7 datasets that we used to study the AMOC slowdown, as well as details of the 15 CMIP5 models used (for more detail, see Table 9.A.1. of ref. [62]).

ARTICLE

The evolutionary history of vertebrate RNA viruses

Mang Shi[1,2,3,8], Xian-Dan Lin[4,8], Xiao Chen[5,8], Jun-Hua Tian[6,8], Liang-Jun Chen[1], Kun Li[1], Wen Wang[1], John-Sebastian Eden[3], Jin-Jin Shen[7], Li Liu[4], Edward C. Holmes[1,2,3] & Yong-Zhen Zhang[1,2]*

Our understanding of the diversity and evolution of vertebrate RNA viruses is largely limited to those found in mammalian and avian hosts and associated with overt disease. Here, using a large-scale meta-transcriptomic approach, we discover 214 vertebrate-associated viruses in reptiles, amphibians, lungfish, ray-finned fish, cartilaginous fish and jawless fish. The newly discovered viruses appear in every family or genus of RNA virus associated with vertebrate infection, including those containing human pathogens such as influenza virus, the *Arenaviridae* and *Filoviridae* families, and have branching orders that broadly reflected the phylogenetic history of their hosts. We establish a long evolutionary history for most groups of vertebrate RNA virus, and support this by evaluating evolutionary timescales using dated orthologous endogenous virus elements. We also identify new vertebrate-specific RNA viruses and genome architectures, and re-evaluate the evolution of vector-borne RNA viruses. In summary, this study reveals diverse virus–host associations across the entire evolutionary history of the vertebrates.

RNA viruses infect a wide range of hosts and contain enormous genetic and phenotypic diversity[1]. Because of their potential effect on public health and the agricultural industries, considerable attention has been directed towards describing the diversity and evolution of RNA viruses associated with vertebrates. Despite an increasingly widespread

surveillance of invertebrate and vertebrate hosts, there are few direct links between invertebrate and vertebrate viruses, and vertebrate viruses tend to form monophyletic groups that are only distantly related to viruses found in invertebrates[1]. Within vertebrates, there has been a marked sampling bias towards mammals and birds[2], even though
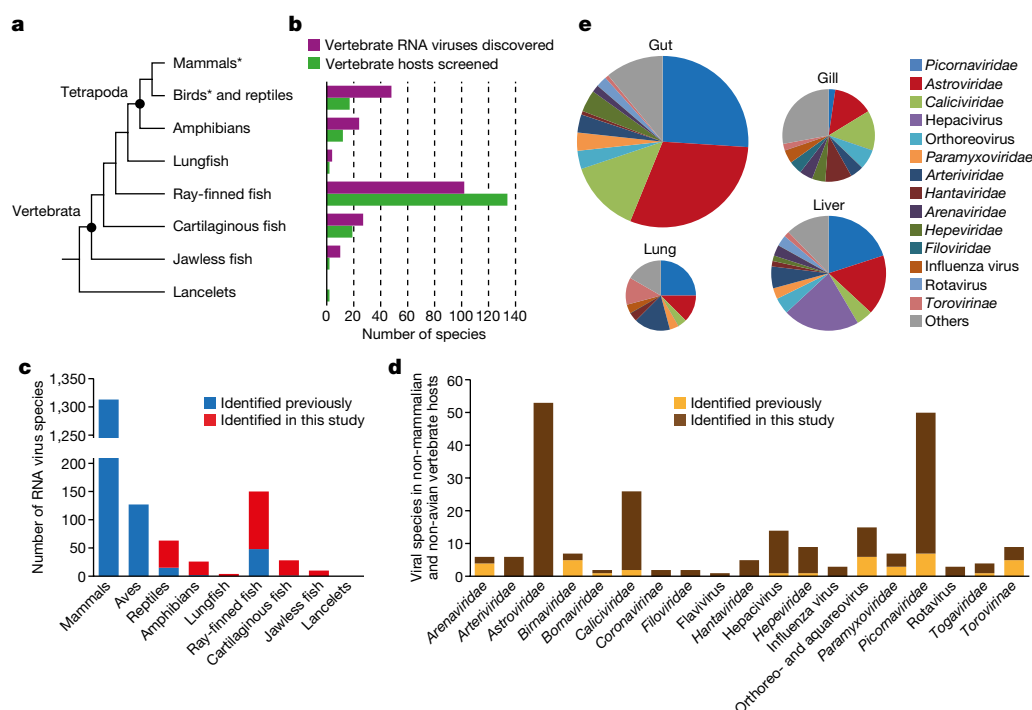

Fig. 1 | Identification of vertebrate-associated viruses in divergent vertebrate host groups. a, Phylogenetic relationships of the vertebrate host classes surveyed here. Asterisks denote hosts not surveyed in this study. **b**, Number of host species surveyed (purple) compared to the number of virus species discovered (green) in each host class. **c**, Number of viral species in each host class. Red and blue represent current and previously discovered RNA viruses, respectively. **d**, Number of non-avian and non-mammalian vertebrate virus species in each vertebrate-associated viral family or genus. Yellow and brown represent current and previously identified RNA viruses, respectively. **e**, Distribution of viruses identified in this study by tissue type.

[1]State Key Laboratory for Infectious Disease Prevention and Control, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China. [2]Shanghai Public Health Clinical Center & Institute of Biomedical Sciences, Fudan University, Shanghai, China. [3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. [4]Wenzhou Center for Disease Control and Prevention, Wenzhou, China. [5]College of Marine Sciences, South China Agricultural University, Guangzhou, China. [6]Wuhan Center for Disease Control and Prevention, Wuhan, China. [7]Yancheng Center for Disease Control and Prevention, Yancheng, China. [8]These authors contributed equally: Mang Shi, Xian-Dan Lin, Xiao Chen, Jun-Hua Tian. *e-mail: zhangyongzhen@shphc.org.cn
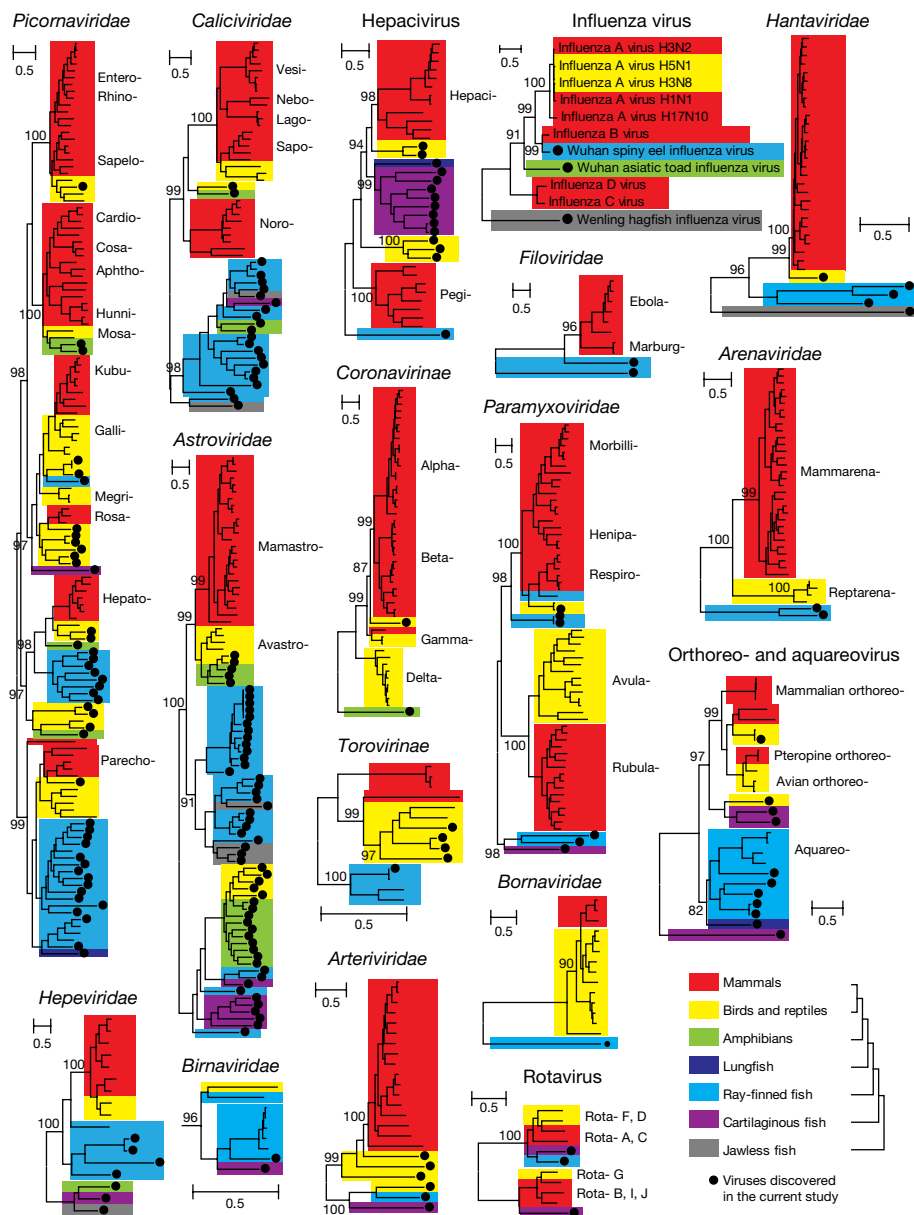
Fig. 2 | Evolutionary history of 17 major vertebrate-specific virus families or genera. Each phylogenetic tree was estimated using a maximum likelihood method, and is rooted using the corresponding broader scale tree that contains both vertebrate and invertebrate viruses (not shown). Within each phylogeny, the viruses newly identified here are marked with solid black circles. Host groups are indicated with different colours; mammals (red), birds and reptiles (yellow), amphibians (green), lungfish (deep blue), ray-finned fish (blue), cartilaginous fish (purple) and jawless fish (grey). The name of the virus family or genus is shown above each phylogeny, and the lower-order virus taxonomy is shown to the right when applicable.

they represent only a small proportion of total vertebrate diversity. Far less is known about those viruses infecting fish, amphibians and reptiles[3], despite their abundance, phenotypic diversity and central role in vertebrate evolution. Notably, the relatively few viruses from fish, amphibians and reptiles documented so far tend to form divergent lineages with respect to known vertebrate RNA viruses[4–7], which in part probably reflects the position of these hosts in the vertebrate phylogeny[8]. However, the extent of viral phylogenetic and genomic diversity in these taxa, their ancestry as well as the relative frequencies of virus–host co-divergence versus cross-species transmission in the evolution of vertebrate RNA viruses remains uncertain[9]. To better understand the origin and evolutionary history of vertebrate viruses, we screened for RNA viruses in a diverse set of species that covered much of the phylogenetic diversity of the vertebrates, including those basal vertebrate lineages in which viruses have only rarely been documented.

Expanding diversity of vertebrate viruses

We performed a large-scale meta-transcriptomics survey of potential vertebrate-associated RNA viruses in more than 186 host species representing the extensive diversity within the phylum Chordata (Fig. 1a, b, Supplementary Table 1). This included animals from the classes Leptocardii (lancelets), Agnatha (jawless fish), Chondrichthyes

(cartilaginous fish), Actinopterygii (ray-finned fish), Sarcopterygii (lungfish), Amphibia (frogs, salamanders and caecilians) and Reptilia (snakes, lizards and turtles). We extracted total RNA from the gut, liver, and lung or gill tissue of these animals, which was then organized into 126 libraries for high-throughput RNA sequencing (Supplementary Table 1). In total, we generated 806 billion bases of sequence reads that were assembled and screened for RNA viruses. Despite the very large number of viruses discovered, we focused on vertebrate-associated viruses, including vertebrate-specific viruses that exhibited relatively close evolutionary relationships to known virus families or genera thought to infect only vertebrate hosts, and 'vector-borne' viruses that are able to infect both vertebrate and invertebrate hosts (Supplementary Table 2). In the resultant phylogenies, the newly discovered viruses either grouped within these families/genera, or fell as immediate sister-groups (Extended Data Figs. 1 and 2). Because the host spectrum of the vertebrate-specific virus families or genera is relatively restricted[2] and generally does not contain viruses associated with other host types[1], we assume that vertebrates were their principle hosts, rather than any eukaryotic or prokaryotic microorganisms also present in the samples. Furthermore, at least 24% of the viruses were recovered from different tissues from the same individual and hence are likely to cause systemic infection (Supplementary Table 2). This gives further
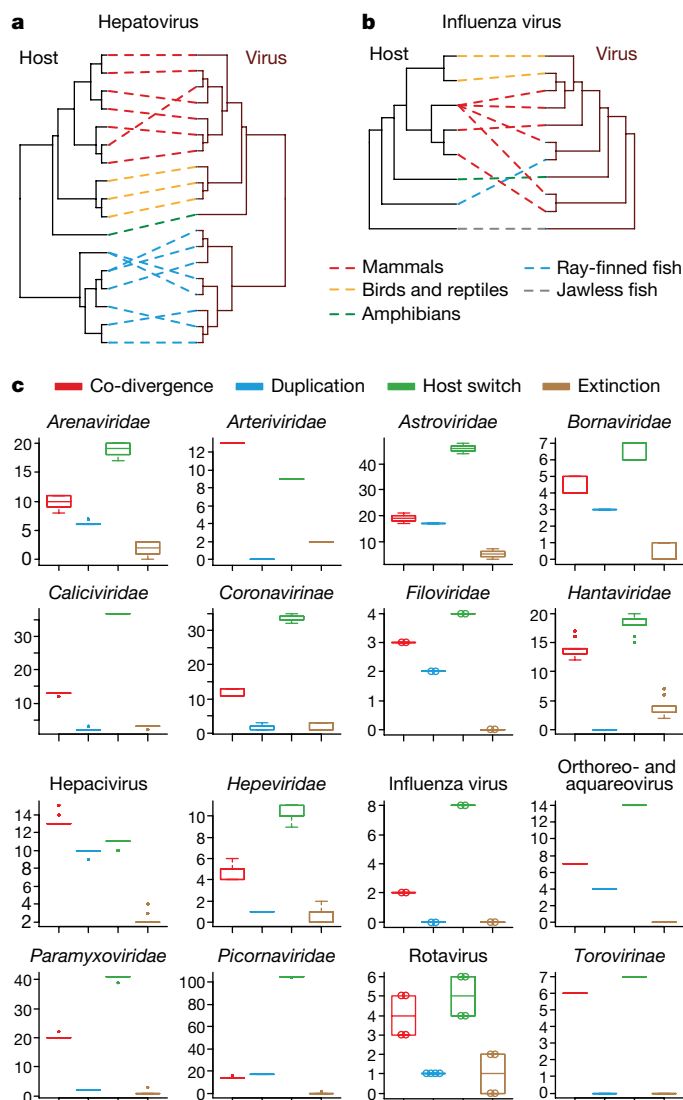
Fig. 3 | Long-term evolutionary relationships between vertebrate hosts and their associated viruses. a, b, Comparisons of hepatovirus (a) and influenza virus (b) phylogenies and their corresponding host phylogenies are presented as examples of virus–host co-divergence and host-switching, respectively. c, Estimation of co-phylogenetic events across the history of vertebrate-associated RNA viruses. Each boxplot illustrates the estimated median (centre line), upper and lower quartiles (box limits), 1.5 × interquartile range (whiskers), and outlier (points) of the co-divergence (red), duplication (blue), host-switching (green) and extinction (brown) events. Data from each estimation (hollow circles) are shown as overlays if there are less than 10 'solutions' provided.

support to a direct association within the vertebrate–host in which they were sampled.

In total, we identified 214 distinctive and previously undescribed putative viral species of vertebrates, of which 196 can be considered vertebrate-specific (Fig. 1b, Supplementary Table 2). Hence, these data reveal that RNA viruses are present in greater numbers and diversity in vertebrates other than birds and mammals than previously realized (Fig. 1c). In particular, it was notable that every vertebrate-specific viral family or genus known to infect mammals and birds is also present in amphibians, reptiles or fish (Fig. 1d). For most of the families or genera, the previously known hosts were either mammals (the *Arteriviridae*, *Filoviridae*, *Hantaviridae* and rubivirus) or mammals, birds and reptiles (*Arenaviridae*, *Astroviridae*, *Bornaviridae*, *Coronavirinae*, influenza virus and rotavirus). This is the first time, to our knowledge, that these viral groups have been identified in fish and/or amphibians (Fig. 1d). Particularly notable was the presence of divergent members of the

Arenaviridae, *Filoviridae* and *Hantaviridae* families in ray-finned fish, suggesting that these previously mammal-dominated groups have relatives in aquatic vertebrates (Fig. 2). Similarly, for those virus groups previously known to contain fish viruses (*Caliciviridae*, *Hepeviridae*, *Paramyxoviridae* and *Picornaviridae*), we were able to greatly expand their genetic diversity, which now covers more phylogenetic space than in their mammalian counterparts (Fig. 2). Of particular note was influenza virus, for which we documented new viruses in jawless fish (hagfish), amphibians (Asiatic toad) and ray-finned fish (spiny eel), with the latter forming a sister-group to human influenza B virus (Fig. 2). Finally, it was notable that the viruses that were newly described in reptiles, amphibians and fish exhibited similar tissue tropisms as their mammalian counterparts[2], which again argues for their antiquity. For example, among the viruses discovered here, those of the Hepacivirus genus were mainly found in the liver, whereas members of the *Picornaviridae*, *Caliciviridae* and *Astroviridae* families dominate in the gut (Fig. 1e).

Long-term virus–host evolutionary relationship

On the basis of the distribution of host taxa on the virus tree, these data also revealed that virus phylogenetic history can mirror that of their hosts over long evolutionary timescales. Most notably, viruses from fish tend to fall basal to viruses in amphibians, reptiles, birds and mammals, reflecting their divergent phylogenetic position within vertebrates (Figs. 2 and 3a). This was supported by the observation that the virus phylogenies exhibited significant clustering by host taxonomy (that is, class), with $P < 0.001$ in the association index (AI)[10] for all family and genus comparisons, with the exception of influenza virus and rotavirus (Table 1). However, despite this overall host clustering, these data also revealed many examples of host-switching during virus evolutionary history. For example, the influenza virus identified in ray-finned fish was the closest relative of mammalian influenza B virus (76% amino acid identity), and the influenza viruses sampled from other tetrapods was more divergent (approximately 30–62% amino acid identity; Fig. 3b). Similarly, the viruses identified in lungfish (in *Picornaviridae*, hepacivirus and aquareovirus; Fig. 2) were more closely related to those from ray-finned fish than to those from tetrapods with whom they share a more recent common ancestor[11].

Table 1 | Phylogenetic test of virus–host association and co-divergence

Virus group	Test of host structure at the level of vertebrate class		Test of virus–host co-divergence		
	Association index ratio*	P value (AI)	Co-divergences	Number of costs	P value (no. of costs)
Arenaviridae	0.0000	<0.001	10–12	27	<0.01
Arteriviridae	0.4960	0.047	13	11	<0.01
Astroviridae	0.0878	<0.001	17–21	68	<0.01
Bornaviridae	0.0020	<0.001	4–5	10	0.05
Caliciviridae	0.2736	<0.001	12–13	42	<0.01
Coronavirinae	0.0834	<0.001	11–13	37	<0.01
Filoviridae	0.0017	<0.001	3	6	0.06
Hantavirus	0.0022	<0.001	12–17	22	<0.01
Hepacivirus	0.0002	<0.001	13–15	23	<0.01
Hepeviridae	0.2935	<0.001	4–6	12	0.13
Influenza virus	0.9173	0.65	2	8	0.8
Orthoreo- and aquareovirus	0.1015	<0.001	7	18	<0.01
Paramyxoviridae	0.0231	<0.001	20–22	44	<0.01
Picornaviridae	0.0294	<0.001	14–15	122	<0.01
Rotavirus	0.9275	0.34	3–5	7	0.52
Torovirinae	0.0072	<0.001	6	7	<0.01

The association index (AI) ratio is calculated as 'observed association index/null association index', in which the null association index is derived from 1,000 tree-tip randomizations. A ratio closer to 0 indicates a stronger host structure. The 'P values (AI) are outcomes from a Bayesian tip-association significance test (BaTS)[10], and derived from 1,000 tree randomizations without adjustment for multiple comparisons. The cost, that is, non-co-divergence, scheme included 'host-switching', 'host duplication', 'host loss' and 'failure to diverge' events, as specified in the model. The 'P values (no. of costs) are outcomes from a co-phylogeny test[12], and are derived from 100 tip-mapping randomizations without adjustment for multiple comparisons.

We next performed a more rigorous co-phylogenetic analysis[12,13] of the resemblance between the virus and host phylogenies at the species level. This revealed significantly more virus–host co-divergence than expected by chance alone (Table 1). However, these data also clearly show that host-switching has been commonplace during the evolutionary history of vertebrate RNA viruses, and is often more frequent than co-divergence across the phylogenies as a whole (Fig. 3c). Aside from phylogenetic position, host-switching is also suggested by the observation that single viruses are occasionally associated with multiple host species or even multiple host orders (such as Beihai fish astrovirus 1 and Wenling fish picornavirus 1; Supplementary Table 2). Collectively, these results suggest that there is a long-term association between the RNA viruses and their vertebrate hosts that stretches many millions of years, but that cross-species transmission has occurred frequently on this background of co-divergence.

To better determine the co-divergence history, we examined the temporal congruence between virus and host evolutionary histories[14,15]. As the large genetic distances between these viruses preclude molecular clock-based studies using heterochronous sequences[16,17], a more profitable approach involves the comparison of exogenous viruses and their endogenous relatives[18]. Previous studies have identified several dating calibration points in the *Filoviridae*[19] and *Bornaviridae*[18,20] families based on the presence of orthologous copies of endogenous virus elements (EVEs) in the genomes of related mammalian species with known times of divergence. Importantly, the viruses newly discovered here in ray-finned fish greatly expand the diversity in both the *Bornaviridae* (Fig. 4a) and *Filoviridae* (Fig. 4b). As a result, both the EVE clades and the calibration points (50 million years (Myr) ago and 30 Myr ago for the *Bornaviridae* and *Filoviridae*, respectively)[18,19] were now deeply nested within the diversity of exogenous viruses, with phylogenetic positions that were relatively distant from the root of the tree. This suggests that both viruses have ancient evolutionary histories that extend well beyond the calibration dates. Although no orthologous EVEs were found in the positive-sense and double-stranded RNA virus families studied here, that their (exogenous) protein sequence divergence was comparable to that of the *Bornaviridae* and *Filoviridae* is also compatible with long evolutionary histories.

Additional vertebrate–associated viruses

We discovered two potentially new groups of vertebrate-associated viruses: one distantly related to the *Astroviridae* and *Potyviridae* families, and another nested within the newly characterized Chuvirus group[21] (Extended Data Fig. 3). Several pieces of evidence support the association of these viruses with vertebrates: (i) they appear in several tissue types (gut, gill and liver), indicative of systemic infection (Extended Data Fig. 3); (ii) a search of the Transcriptome Shotgun Assembly (TSA) sequence database revealed that related viral sequences were found only in vertebrate transcriptomes, again involving several tissue types (Extended Data Fig. 3); and (iii) in the case of the vertebrate-associated chuviruses, EVEs were found in the genomes of several species of ray-finned fish (Extended Data Fig. 3).

In addition to the vertebrate-specific viruses, we discovered viruses in amphibians, fish and reptiles from genera that have previously been associated with vector-borne virus transmission, most notably alphaviruses, dimarhabdoviruses and flaviviruses. Among these, Wenzhou shark flavivirus is the first member of the Flavivirus genus identified in cartilaginous fish, and was found in all the tissue types analysed compatible with a systemic infection (Supplementary Table 2). In the phylogeny, Wenzhou shark flavivirus falls basal to the 'classic' vector-borne and insect-specific flaviviruses, and was more closely related to Tamana bat virus that has no known vector species (Extended Data Fig. 4). Similarly, in the case of the alphaviruses and dimarhabdoviruses, the fish viruses discovered here clustered with other fish viruses reported previously to form lineage(s) basal to those associated with vector-borne viruses (Extended Data Fig. 4). This complex mix of vectored and non-vectored viruses, with clear cases of the secondary loss of vector-borne transmission (Extended Data Fig. 4), raises the question
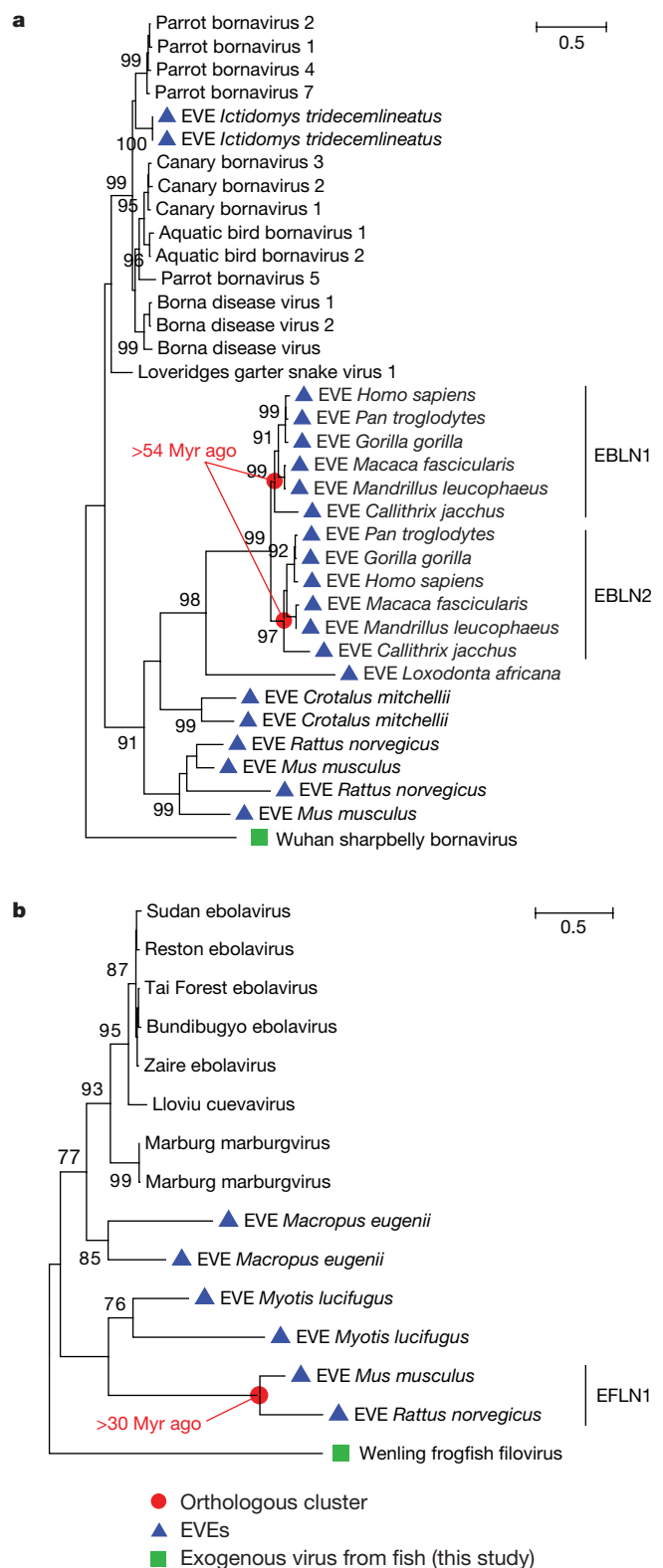
Fig. 4 | Evaluating the timescale of vertebrate virus evolution using EVEs. a, b, Phylogenies were based on the exogenous and endogenous nucleoproteins for the *Bornaviridae* (**a**) and *Filoviridae* (**b**) families. Within the trees, endogenous virus elements (EVEs) are highlighted with blue triangles, and the (divergent) exogenous viruses discovered in fish are highlighted with green squares. The nodes that represent orthologous clusters are highlighted with red circles, their associated divergence times are shown next to the nodes, and their corresponding names are given to the right of the phylogeny.

Fig. 5 | Evolution of vertebrate-associated virus genomes. Representative genomes from 12 vertebrate-associated virus families or genera are shown. The regions that encode ma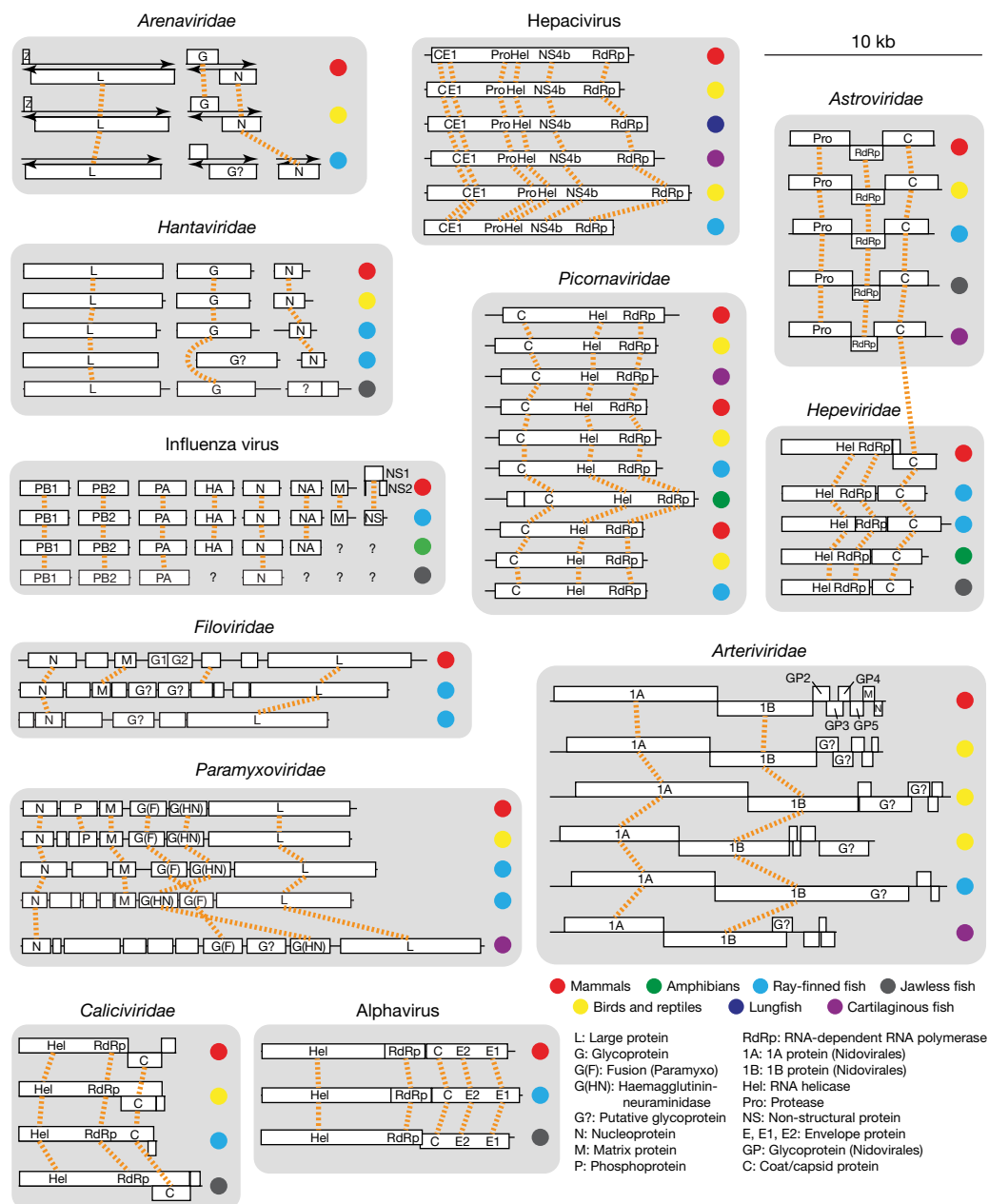jor functional proteins or protein domains are labelled on each of the genomes. Homologous regions within or between viral families are connected by orange dotted lines. Host associations are labelled to the right of each genome using solid circles with different colours. The orientation of the positive-sense genomes are shown from 5′ to 3′, those of negative-sense genomes are from 3′ to 5′, and those of ambisense genomes (that is, arenaviruses) are indicated using arrows. More detailed depictions of genome evolution are presented in Extended Data Figs. 5 and 6.

of whether some of the vector-borne viruses were ultimately derived from vertebrate-specific or vector-specific viruses, or if the ability to infect both arthropods and vertebrates is the ancestral phenotype[22].

Genome evolution of vertebrate RNA viruses

The annotation of the virus genomes newly documented here showed a wider variety of genome architectures for vertebrate virus families or genera than previously observed[2], some of which may represent the ancestral types in the evolutionary history of these viruses (Fig. 5). Although the structures of these vertebrate virus genomes were more conserved than those of invertebrates[1,6,21,23], they still exhibited extensive variation, including genome length (hepacivirus), the organization of open reading frames (*Caliciviridae*), the complete re-configuration of the genome downstream of the non-structural genes (*Arteriviridae*), changes

in the order and number of glycoproteins (*Paramyxoviridae*), inter-species re-assortment involving the M segment (hantavirus), inter-family recombination involving the capsid protein (*Astroviridae* and *Hepeviridae*)[24] and changes in segment numbers in the *Arenaviridae* family (Fig. 5, Extended Data Figs. 5 and 6). The latter is particularly interesting as the *Arenaviridae* were traditionally thought to be a family of bi-segmented negative-sense RNA viruses[2]. However, we discovered two arenavirus species in fish with genomes comprising three segments, similar to that of the divergent relative of the *Arenaviridae* family found in arthropods[21], and suggesting that there was a decrease in segment numbers from three to two (Extended Data Fig. 5). If so, this represents an important example of a reduction in segment number without a corresponding loss in gene content, hence compatible with segment merging.

Discussion

Despite a combination of rapid evolution and frequent host-switching, our large-scale analysis of virus diversity in previously undersampled hosts suggests that RNA viruses in vertebrates tend to broadly follow the evolutionary history of their hosts that began in the ocean and extends for hundreds of millions of years. These results, which apply to most of the vertebrate RNA virus families or genera, are in accord with recent analyses of viral evolution using palaeovirological data[18–20,25,26], and demonstrate the importance of conducting widespread taxonomic surveys of virus diversity when trying to reveal evolutionary history. These results also have broader implications for our understanding of virus evolution. In particular, it is clearly simplistic and perhaps erroneous to identify a specific host group as ancestral to another given that our sampling of RNA virus diversity is still so very limited. For example, on current data we suggest that it is premature to conclude that vertebrate RNA viruses necessarily originated in mosquitoes/ticks, since it is possible that the evolution of specific virus families may have followed that of the metazoans over an even longer period of co-divergence. In summary, our study reveals long-term virus–host relationships for each vertebrate-associated virus family that extend over geological timescales, further illustrating the ancient history of RNA viruses.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0012-7.

1. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
2. King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. *Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic, Amsterdam, 2012).
3. Essbauer, S. & Ahne, W. Viruses of lower vertebrates. *J. Vet. Med. B Infect. Dis. Vet. Public Health* **48**, 403–475 (2001).
4. Batts, W., Yun, S., Hedrick, R. & Winton, J. A novel member of the family Hepeviridae from cutthroat trout (*Oncorhynchus clarkii*). *Virus Res.* **158**, 116–123 (2011).
5. Mikalsen, A. B. et al. Characterization of a novel calicivirus causing systemic infection in atlantic salmon (*Salmo salar* L.): proposal for a new genus of caliciviridae. *PLoS ONE* **9**, e107132 (2014).
6. Shi, M. et al. Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the *Flaviviridae* and related viruses. *J. Virol.* **90**, 659–669 (2015).
7. Stenglein, M. D. et al. Identification, characterization, and *in vitro* culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease. *MBio* **3**, e00180–12 (2012).
8. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol* **32**, 835–845 (2015).
9. Dill, J. A. et al. Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *J. Virol.* **90**, 7920–7933 (2016).
10. Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E. & Simmonds, P. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75**, 11686–11699 (2001).
11. Brinkmann, H., Venkatesh, B., Brenner, S. & Meyer, A. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc. Natl Acad. Sci. USA* **101**, 4900–4905 (2004).
12. Conow, C., Fielder, D., Ovadia, Y. & Libeskind-Hadas, R. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol. Biol.* **5**, 16 (2010).
13. Geoghegan, J. L., Duchêne, S. & Holmes, E. C. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* **13**, e1006215 (2017).
14. Charleston, M. A. & Robertson, D. L. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol* **51**, 528–535 (2002).
15. de Vienne, D. M. et al. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytol.* **198**, 347–385 (2013).
16. Wertheim, J. O. & Kosakovsky Pond, S. L. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol* **28**, 3355–3365 (2011).
17. Zhang, Y. Z. & Holmes, E. C. What is the time-scale of hantavirus evolution? *Infect. Genet. Evol.* **25**, 144–145 (2014).
18. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191 (2010).
19. Taylor, D. J., Leach, R. W. & Bruenn, J. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol. Biol.* **10**, 193 (2010).
20. Horie, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
21. Li, C. X. et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4**, e05378 (2015).
22. Longdon, B. et al. The evolution, diversity, and host associations of rhabdoviruses. *Virus Evol.* **1**, vev014 (2015).
23. Qin, X. C. et al. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. *Proc. Natl Acad. Sci. USA* **111**, 6744–6749 (2014).
24. Kelly, A. G., Netzler, N. E. & White, P. A. Ancient recombination events and the origins of hepatitis E virus. *BMC Evol. Biol.* **16**, 210 (2016).
25. Han, G. Z. & Worobey, M. An endogenous foamy-like viral element in the coelacanth genome. *PLoS Pathog.* **8**, e1002790 (2012).
26. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).

Reviewer information *Nature* thanks A. Rambaut and M. Worobey for their contribution to the peer review of this work.

Author contributions M.S. and Y.-Z.Z. conceived and designed the study. M.S., X.-D.L., X.C., J.-H.T., K.L., L.-J.C., J.-J.S., L.L. and Y.-Z.Z. organized field work, and collected samples. M.S., X.-D.L., X.C., J.-H.T., K.L., L.-J.C., W.W., J.-J.S., L.L. and Y.-Z.Z. performed experiments. M.S., J.-S.E., E.C.H. and Y.-Z.Z. analysed data. M.S., E.C.H. and Y.-Z.Z. wrote the paper with input from all authors. Y.-Z.Z. led the study.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0012-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0012-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.-Z.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample collection. The goal of this study was to survey animal species that were representative of biological diversity within the phylum Chordata, and that have only rarely been analysed for the presence of RNA viruses. Accordingly, we focused on amphibians, reptiles and fish rather than birds and mammals that have been studied in far greater detail (Fig. 1d). We also targeted species distributed at diverse locations across the vertebrate phylogeny (Fig. 1a), although those species associated with most basal vertebrate lineages are often rare. For each species we sampled 1–24 individuals to represent a population. No statistical methods were used to predetermine sample size. The procedures for sampling and sample processing were approved by the ethics committee of the National Institute for Communicable Disease Control and Prevention of the Chinese CDC.

In total, we sampled two species from the subphylum Cephalochordata (that is, lancelets), with the remainder from the subphylum Vertebrata (Supplementary Table 1, Fig. 1a). Within Vertebrata, we sampled two species each from the classes Agnatha (that is, jawless fish) and Sarcopterygii (that is, lungfish), as these are relatively rare. Most of our aquatic samples were from the classes Chondrichthyes (cartilaginous fish), from which we sampled 19 species, and Actinopterygii (bony fish), from which we sampled more than 130 species across 20 orders (Supplementary Table 1). With respect to land tetrapods, we sampled 12 species from the class Amphibia, including the orders Aura, Caudata and Gymnophiona, and 17 species from the class Reptilia, including the orders Testudines and Squamata (Supplementary Table 1).

With the exception of lungfish samples, which were obtained from Nigeria (*Protopterus annectens*) and Chile (*Lepidosiren* sp.), respectively, all other samples were collected in China (Supplementary Table 1). The marine species were sampled from the South China Sea, East China Sea and Yellow Sea, mostly from fishing vessels. The samples were kept at −20 °C on the boat before being transferred to −80 °C for storage. The remaining marine samples were either collected frozen from the returned ships at the dock, or purchased alive from local fisherman at nearby markets. The freshwater fish samples were captured alive using fishing rods or nets from rivers and lakes in Hubei, Heilongjiang and Guangdong provinces. The reptile and amphibian samples were caught by field biologists from a wide range of geographic locations, including Fujiang, Guangdong, Guangxi, Xinjiang and Zhejiang provinces.

For most of the animal samples, three types of internal organs were harvested, comprising the gut, liver and gill for jawless, cartilaginous, and ray-finned fish, and gut, liver and lung for amphibians and reptiles (Supplementary Table 1). For lungfish, all four types of tissue (that is, gut, liver, lung and gill) were obtained. For lancelets, the entire individual was used owing to their small body size. All specimens were stored at −80 °C for later RNA extraction.

Host species information was initially identified by experienced field biologists on capture based on morphological traits, and was later confirmed by sequencing and analysing the partial cytochrome *c* oxidase (COI) gene from each sample (approximately 600–700 nucleotides near 5′ of the gene).

RNA library construction and sequencing. RNA was extracted from individual animal specimens. For the initial screening of viruses, aliquots of RNA from several (that is, from 13 to 62) individuals of a particular taxonomic group or multiple taxonomic groups were pooled for library preparation and sequencing (Supplementary Table 1). After determining the presence of a specific virus, a subset of the initial pool or the individual un-pooled RNA extractions was subject to library construction and sequencing to obtain better genome coverage (Supplementary Table 1).

For each RNA extraction, we first transferred approximately 30 mg from the specimen to 500–700 μl standard, sterile, RNA and DNA-free 1× PBS solution (GIBCO). The tissue was then homogenized in the PBS solution using the Mixer mill MM400 (Restsch). Total RNA was extracted from the homogenates using TRIzol LS reagent (Invitrogen) and subsequently purified using RNeasy Plus Mini Kit (Qiagen). Aliquots of the resultant RNA solutions were then pooled in equal quantity. The quality of the pooled RNA was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies) before library construction and sequencing.

The TruSeq total RNA Library Preparation protocol (Illumina) was used for all library preparations. Ribosomal (r) RNA was removed using the Ribo-Zero Gold (Epidemiology) Kit (Illumina) for most of the libraries, with the exception of LXMC-PolyA and XYHYMC-PolyA for which poly(A) enrichment was used (Supplementary Table 1). The average fragmentation size for these libraries was either 200 bp or 300 bp. Accordingly, 100 bp and 150 bp paired-end sequencing of the RNA libraries were performed on the Hiseq 2500 and HiSeq 4000 platforms (Illumina), respectively. All library preparation and sequencing was carried out by BGI Tech (Shenzhen).

RNA virus discovery. For each library, sequencing reads were first adaptor- and quality-trimmed using the Trimmotic program[27] with the following options: SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, MINLEN:25. The remaining reads were assembled de novo using the Trinity program (version 2.1)[28] with

default parameter settings. To identify viral contigs, the assembled contigs were compared (using blastx) against the database comprising reference RNA virus proteins downloaded from GenBank. The *E*-value cut-off for these comparisons was set at 1×10^{-5}. To eliminate false positives, these putative viral contigs were compared against the entire non-redundant nucleotide and protein database. The remaining contigs with unassembled overlaps were merged to form longer viral contigs using the SeqMan program implemented in the Lasergene software package (version 7.1, DNAstar).

Among all the virus contigs discovered, those likely to be associated with vertebrates (that is, vertebrate-specific viruses and vector-borne vertebrate viruses) were initially identified based on a closer relationship to established vertebrate-associated viruses than to other taxa in a Blast analysis (that is, known vertebrate-associated viral families/genera were the top blast hits). This relationship was later confirmed by more detailed phylogenetic analyses including viruses representative of both vertebrates and a wider variety of non-vertebrate organisms[1,6,21] (Extended Data Figs. 1 and 2).

For the vertebrate-associated viruses, we determined which samples contained the viruses and hence its potential host(s) using PCR with reverse transcription (RT–PCR) and sequencing. Accordingly, for each virus, we designed 2–3 pairs of primers based on the viral contigs and screened all the unpooled RNA extractions of the corresponding library. The target PCR products were then validated by Sanger sequencing.

Gaps in incomplete vertebrate-associated virus genomes were filled by either RT–PCR or by re-sequencing (using the meta-transcriptomics approach described above) on the individual RNA samples that contained the target virus. Genome termini were determined by RNA circularization as previously described[23], or by using the 5′/3′ RACE kits (TaKaRa). Confirmation of most of the viral genome sequences was performed by read mapping using Bowtie2[29], with the final majority consensus sequences determined from the final assembly of mapped reads using Geneious v.8[30]. For virus species with multiple variants in the same pool, we performed meta-transcriptomics or RT–PCR and Sanger sequencing of the entire genome from the individual positive sample. To exclude the possibility that these contigs belonged to expressed EVEs (see below), we used PCR and Sanger sequencing to examine the DNA extracted from the homogenates of the corresponding samples.

Searching existing databases for vertebrate viruses. To discover more vertebrate-associated viruses and hence enrich our dataset, we downloaded the entire Transcriptome Shotgun Assembly (TSA) sequence database which was then used as query to search against the virus protein database as previously described. Because not all transcriptome sequences have a TSA (assembled) entry, we also examined reads deposited in the Sequence Read Archive (SRA) database. We targeted basal vertebrate taxa with inadequate or limited sampling, including lancelets (NCBI taxonomy ID: 7736), jawless vertebrates (NCBI taxonomy ID: 1476529), cartilaginous fish (NCBI taxonomy ID: 7777), and lungfish (NCBI taxonomy ID: 7878). These reads were assembled using Trinity and compared against the virus protein database as described above. Unfortunately, no vertebrate-associated viruses were found in these read archives.

To reveal viruses that may have infected vertebrates in the evolutionary past, we searched within the vertebrate genomes for EVEs that were relatively closely related to the viruses discovered in this study, especially those that did not belong to any established vertebrate clade. Accordingly, we first downloaded all the assembled genome sequences within the taxonomic group Vertebrata (NCBI taxonomy ID: 7742) from the NCBI genome FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). We then compared the translated viral protein sequences discovered in this study against the all assembled vertebrate genomes using the tblastn program, with an *E*-value cut-off set at 1×10^{-20}. For each potential EVE, the query process was reversed to determine their phylogenetic positions. The alignment of EVEs and exogenous viruses was checked manually to exclude false-positives.

Virus genome characterization. For newly identified virus genomes, the prediction of the potential open reading frames (ORFs) was based on those from the related reference virus genomes. The annotation of ORFs was first based on comparisons against the Conserved Domain Database (CDD) and then against the non-redundant protein database. The remaining proteins were characterized by predicting their primary protein structure using the programs NetNGlyc, SignalP, and TMHMM (http://www.cbs.dtu.dk/services/). For example, the divergent glycoprotein genes of negative-sense RNA families were identified based on the presence of (i) an N-terminal signal peptide, (ii) a mid-point or C-terminal transmembrane domain, and (iii) putative *N*-linked glycosylation sites. Finally, the sequencing depth of each viral genome within the library was estimated based on the percentage of total reads that mapped to the target genome.

In the case of segmented viruses, most of the non-RdRp segments were recovered by homology comparisons. However, divergent members of the families *Hantaviridae* and *Arenaviridae* had glycoproteins that lacked clear homology with those of other family members. To look for these segments we first annotated

all contigs that were of similar sequencing depths by comparing them to the nr database. This removed most sequences of host origin. For the remaining contigs, we examined (i) the potential glycoprotein structure (that is, signal peptide, transmembrane domains and glycosylation sites), (ii) the presence of inverted complementary genome termini that are the same to those of other segments, (iii) whether all the segments were found in the same samples and (iv) whether its closest relative contained the related segment. Only when all four criteria were satisfied did we conclude that these segments most likely belonged to the same virus.

Inferring virus evolutionary history. We examined the phylogenetic relationship among these viruses at two levels: (i) an overall evolutionary history that placed the vertebrate-associated viruses in the context of viruses sampled from other hosts, and (ii) family/genus specific phylogenies that provide a more detailed depiction of the evolutionary relationships within each of the vertebrate-associated virus families/genera. At the family/genus level, we included as background all reference virus replicase sequences (that is, RNA-dependent RNA polymerase; RdRp) as well as replicases from non-reference viruses that occupied a unique phylogenetic position and which had an established host association. At the overall level, we included viral replicases representative of a broader phylogenetic diversity[1,6,21] in addition to those used in the family/genus level phylogenies.

For each dataset, the virus replicases were aligned using the E-INS-i algorithm implemented in the program MAFFT (version 7)[31], with all ambiguously aligned regions were subsequently removed using TrimAl (version 1.2)[32]. The best-fit model of amino acid substitution in each dataset was determined using ProtTest (version 3.4)[33]. Phylogenetic trees were then inferred using the maximum likelihood method implemented in PhyML (version 3.0)[34], using the best-fit substitution model and Subtree Pruning and Regrafting (SPR) branch-swapping. Support for specific nodes on the trees was assessed using an approximate likelihood ratio test (aLRT) with the Shimodaira–Hasegawa-like procedure. In addition, phylogenetic trees were inferred using the Bayesian method implemented in the program MrBayes v.3.2[35], using the same amino acid substitution models. Because the tree topologies generated by the two programs were largely identical, only maximum likelihood phylogenies are shown here.

Examining virus–host evolutionary relationships. We used the BaTS (Bayesian tip-association significance testing) program[36] to test whether viruses cluster more strongly with particular host taxonomic groups than expected by chance alone. This analysis considered host phylogenetic structure at the level of vertebrate class: that is, mammals, reptiles and birds, amphibians, lungfish, bony fish, cartilaginous fish, and jawless fish. Accordingly, we estimated the association index[10] within BaTS to determine the strength of the association between virus phylogeny and host class. This was then compared to a null distribution generated using 1,000 replicates of state randomization across a credible set of trees generated by MrBayes as described above.

To examine the extent of virus-host co-divergence in each vertebrate-specific virus family/genus, we performed event-based co-phylogenetic reconstructions using the Jane program (version 4)[12]. The virus phylogenies were based on the family/genus level phylogenies estimated here, from which we removed those with no host information. All 'generalist' viruses (that is, those that infect more than three species of hosts) were included in the analyses as unresolved parallel lineages. The corresponding host topologies were obtained from both the TIMETREE website (http://www.timetree.org/) and a previous phylogeny of bony fish[37]. The 'cost' scheme for analyses in Jane was set as follows: co-divergence = 0, duplication = 1, host switch = 1, loss = 1, failure to diverge = 1. The number of generations and the population size were both set to 100. The significance of co-divergence was derived by comparing the estimated costs to null distributions calculated from 100 randomizations of host tip mapping.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All sequence reads generated in this study are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA418053 (Supplementary Table 1). All viral sequences generated in this study have been deposited in GenBank under the accession numbers MG599863–MG600130 (Supplementary Table 2). All virus nucleotide sequences (fasta format), the unaligned and the aligned data set used in the phylogenetic analyses (fasta format), as well as the phylogenetic trees (newick and MEGA5 mts format), are available at the Figshare website at: https://Figshare.com/articles/The_evolutionary_history_of_vertebrate_RNA_viruses/5405620.

27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
28. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
32. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
33. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
34. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol* **52**, 696–704 (2003).
35. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
36. Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).
37. Betancur-R, R. et al. The tree of life and a new classification of bony fishes. *PLoS Curr.* **5**, https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288 (2013).
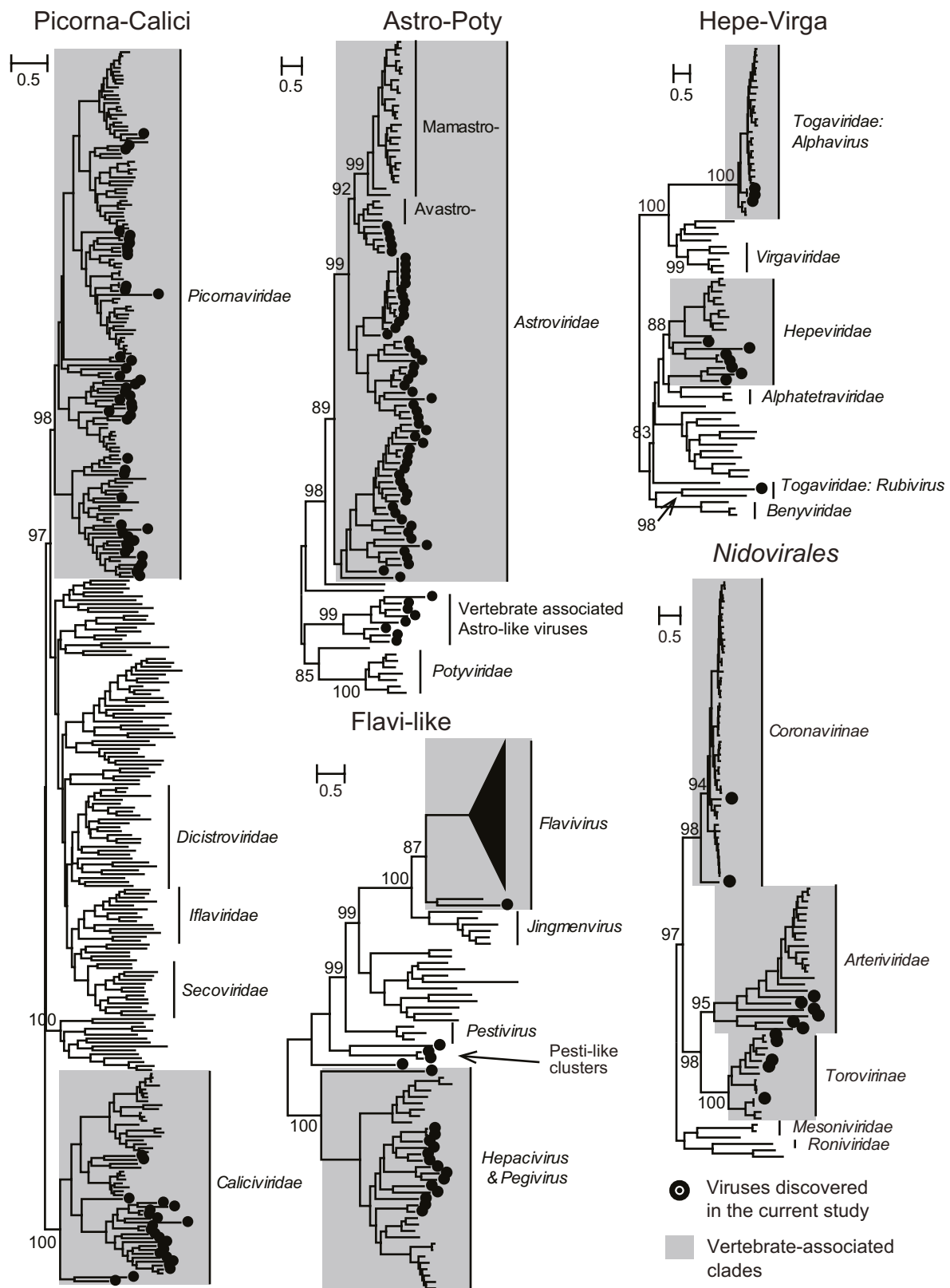
Extended Data Fig. 1 | Phylogenetic positions of vertebrate-associated positive-sense and double-stranded RNA viruses within the broader diversity of RNA viruses. Phylogenies were estimated using a maximum likelihood method and midpoint-rooted for clarity only. Viruses discovered here are labelled with solid black circles. The name of the major clade (phylogeny) is shown at the top of each tree, and taxonomic names are shown to the right. The vertebrate associated virus diversity is shaded in grey. All horizontal branch lengths are scaled to the number of amino acid substitutions per site.
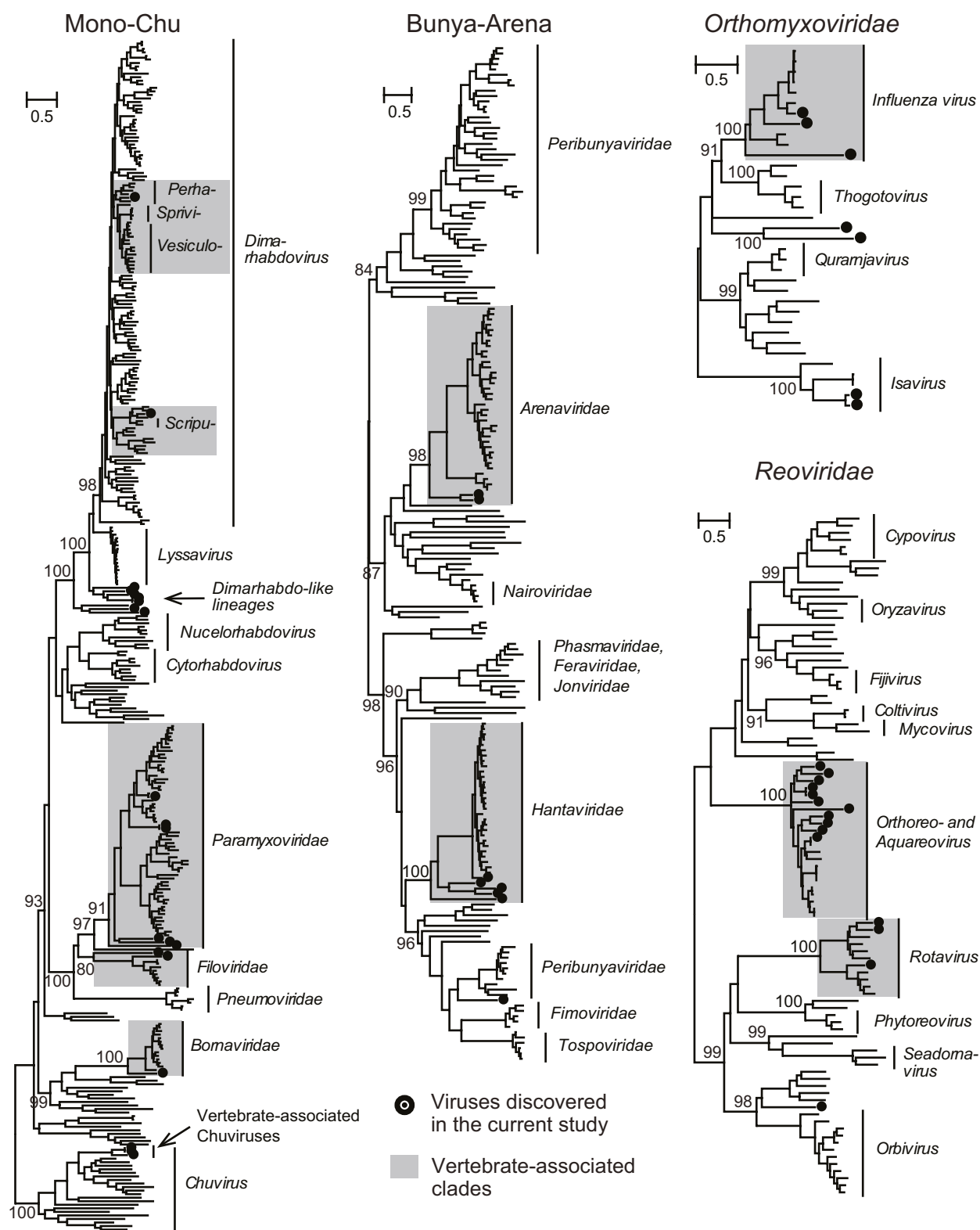
Extended Data Fig. 2 | Phylogenetic positions of vertebrate-associated negative-sense RNA viruses within the broader diversity of RNA viruses. Phylogenies were estimated using a maximum likelihood method and midpoint-rooted for clarity only. Viruses discovered here ar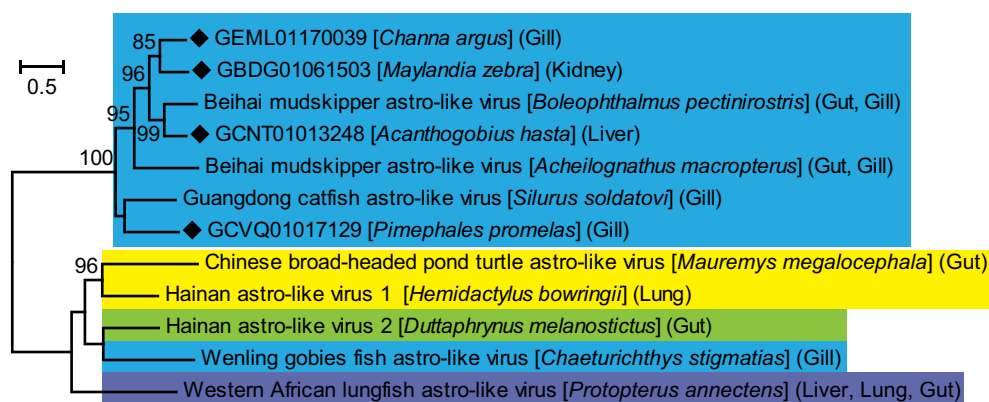e labelled with solid black circles. The name of the major clade (phylogeny) is shown at the top of each tree, and taxonomic names are shown to the right. The vertebrate associated virus diversity is shaded in grey. All horizontal branch lengths are scaled to the number of amino acid substitutions per site.
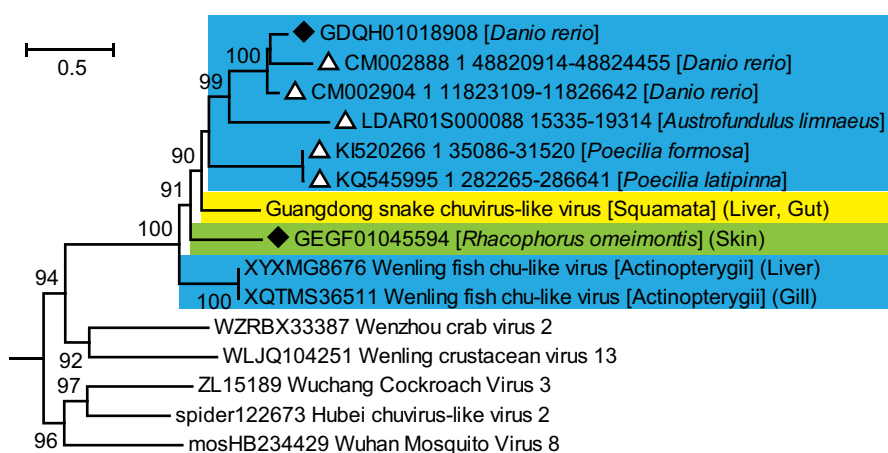
Vertebrate-associated astro-like viruses

Vertebrate-associated Chuviruses

◆ Sequences recovered from the Transcriptome Shotgun Assembly (TSA) Database

△ Sequences recovered from the Whole-Genome Shotgun Contigs (WGS) Database
(i.e. endogenous virus elements)

Birds and reptiles   **Amphibians**   **Ray-finned fish**   **Lungfish**

Extended Data Fig. 3 | The phylogenies of potentially new families of vertebrate-associated viruses. Viruses identified from vertebrate hosts are shaded with different colours. Sequences recovered from the Transcriptome Shotgun Assembly (TSA) database are marked with solid black diamonds, while those recovered from the Whole-Genome Shotgun (WGS) contigs database (that is, endogenous virus elements) are marked with open triangles. For vertebrate viruses, the relevant taxonomic and tissue information is provided in the sequence names.
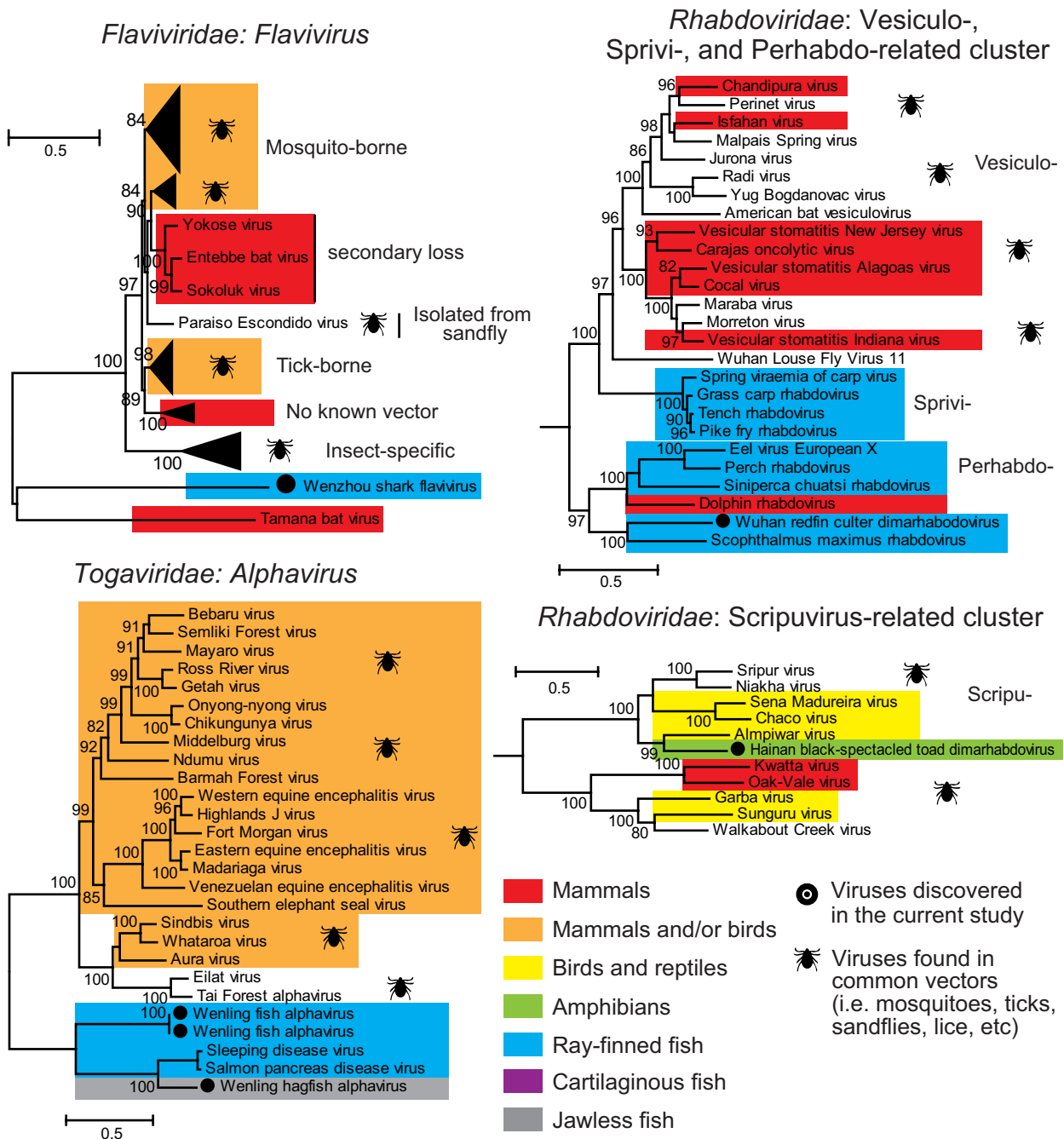
Extended Data Fig. 4 | Evolutionary history of four groups of vector-borne RNA virus. Each phylogenetic tree was estimated using a maximum likelihood method. Within each phylogeny, the viruses newly identified here are marked with solid black circles, the vertebrate host groups are indicated by different colours, and the vector symbol is shown next to viruses known to be transmitted by vectors. The name of the virus family or genus is shown at the top of each phylogeny, and the lower level virus taxonomic names are shown to the right.

Extended Data Fig. 5 | Evolution of vertebrate-associated negative-sense RNA virus genomes. Representative genomes from negative-sense RNA virus families/genera are shown. The regions that encode major functional proteins or protein domains are labelled on each of the genomes. Homologous regions within each family are connected with orange dotted lines. Schematic phylogenetic relationships are shown next to the genomes diagrams. Coverage plots are shown underneath novel genome structures. Reverse-complementary sequences are shown for negative-sense RNA viruses with complete termini. A Sanger sequencing chromatogram is shown at a GC-rich hairpin-forming region of the Wenling frogfish arenavirus 2 genome, in which the coverage drops substantially. Host associations are labelled to the right of tree using solid circles with different colours. Host associations and abbreviation of functional domains are described at the bottom of the figure.
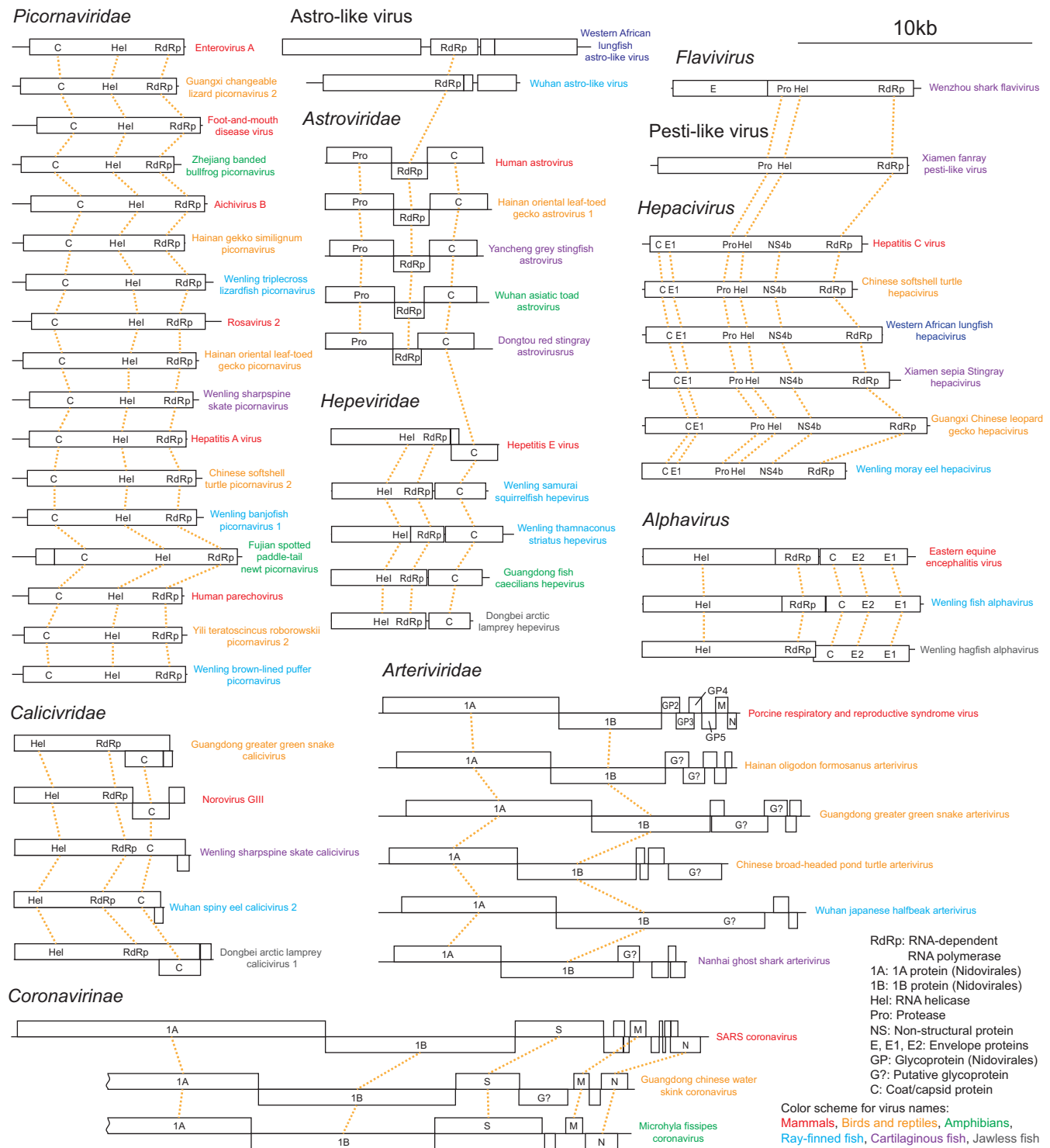
Extended Data Fig. 6 | Evolution of vertebrate-associated positive-sense RNA virus genomes. Representative genomes from positive-sense RNA virus families or genera are shown. The regions that encode major functional proteins or protein domains are labelled on each of the genomes. Homologous regions within or between viral families are connected by orange dotted lines. Host associations are reflected in the colour of the virus names. Host association colour schemes and the abbreviations of functional domains are described at the bottom of the figure.

ARTICLE

Cryo-EM structure of the *Blastochloris viridis* LH1–RC complex at 2.9 Å

Pu Qian[1]*, C. Alistair Siebert[2], Peiyi Wang[3], Daniel P. Canniffe[1] & C. Neil Hunter[1]*

The light–harvesting 1–reaction centre (LH1–RC) complex is a key functional component of bacterial photosynthesis. Here we present a 2.9 Å resolution cryo-electron microscopy structure of the bacteriochlorophyll *b*-based LH1–RC complex from *Blastochloris viridis* that reveals the structural basis for absorption of infrared light and the molecular mechanism of quinone migration across the LH1 complex. The triple-ring LH1 complex comprises a circular array of 17 β-polypeptides sandwiched between 17 α- and 16 γ-polypeptides. Tight packing of the γ-apoproteins between β-polypeptides collectively interlocks and stabilizes the LH1 structure; this, together with the short Mg–Mg distances of bacteriochlorophyll *b* pairs, contributes to the large redshift of bacteriochlorophyll *b* absorption. The 'missing' 17th γ-polypeptide creates a pore in the LH1 ring, and an adjacent binding pocket provides a folding template for a quinone, Q_P, which adopts a compact, export-ready conformation before passage through the pore and eventual diffusion to the cytochrome bc_1 complex.

Photosynthesis provides the energy for almost all life on Earth. In the early stages of photosynthesis, light-harvesting complexes absorb solar energy, which is transferred to a membrane-bound RC, where a charge separation initiates the eventual formation of a reduced electron acceptor[1–3]. The basic functional unit in purple phototrophic bacteria is LH1–RC, the complex of LH1 and the RC, in which the RC is surrounded by a ring-like oligomeric assembly of LH1 α- and β-heterodimers that bind bacteriochlorophyll (BChl) and carotenoid pigments. LH1–RC complexes in different species exhibit a variety of architectures: 16 LH1 α–β pairs completely encircle the RC in the *Thermochromatium tepidum*[4] and *Rhodospirillum rubrum*[5] complexes, whereas in *Rhodopseudomonas palustris* the RC is encircled by an open LH1 ring consisting of 15 α–β pairs and a W polypeptide[6]. The *Rhodobacter sphaeroides* complex has a dimeric core[7], in which each monomer has 14 α–β pairs associated with one RC; two monomers associate through two PufX polypeptides to form an S-shaped LH1 ring[8,9].

A high level of structural detail is required to account for the ability of LH1–RC complexes to absorb solar energy within a specific spectral range and to drive the formation of a quinol, which must traverse the confines of the LH1 ring encircling the RC. We identified the LH1–RC complex from *Blastochloris (Blc.) viridis* as a suitable target for a high-resolution structural study because it possesses unique architectural and spectroscopic features. Notably, the RC in the *Blc. viridis* photosynthetic complex yielded the first reported structure of a membrane protein complex[10], but electron microscopy has provided only low-resolution structures for the complete LH1–RC complex[11,12]. This complex accommodates BChl *b* rather than BChl *a*, and absorbs in the infrared at 1,015 nm, making it one of the most redshifted photosynthetic complexes described to date and one proposed as the basis for re-engineered photosynthesis[13]. There is currently no known structural basis for this unusual in vivo absorption, which represents one of the largest redshifts observed in a photosynthetic pigment–protein complex, 220 nm from the 795 nm absorption maximum of BChl *b* in methanol. This property could be related to the composition of the *Blc. viridis* LH1 complex, which contains α-, β- and γ-polypeptides, but the position and function of the γ-subunit within the LH1

complex remain poorly understood. The *Blc. viridis* LH1 contains rare 1,2-dihydro-derivatives of neurosporene and lycopene as major carotenoids[14–16]. The LH1–RC complex forms extensive arrays in the lamellar membranes of *Blc. viridis*[17–20], which have been hypothesized to consist of closed 16-membered LH1 rings that completely encircle each RC[20]. However, such an arrangement of LH1 subunits represents a potential obstacle for quinol export from the RC to the external quinone pool in the membrane and eventual reduction of the cytochrome bc_1 complex. Here, we report a 3D cryo-electron microscopy (cryo-EM) structure at 2.9 Å resolution of this BChl *b*-based photosynthetic complex from *Blc. viridis*. Analysis of this structure shows how γ-apoproteins influence the large redshift observed in the BChl *b*-Q_y-absorption band, reveals the position of the internal quinone channel, and identifies a third quinone binding site that prepares quinol for export through the pore in the LH1 ring.

Overall structure of the LH1–RC complex

Extended Data Fig. 1 shows the absorption spectra of native photosynthetic membranes and LH1–RC complexes purified from *Blc. viridis*. The absorption maximum at 1,015 nm is ascribed to the Q_y band of BChl *b* in the LH1 complex, which is slightly blueshifted to 1,008 nm after detergent solubilization and purification. Following vitrification of monodisperse complexes, we recorded 6,472 cryo-EM movies, from which 267,726 particles were picked manually for reference-free 2D classification. Further processing yielded a final resolution of 2.9 Å, enabling compilation of a colour-coded electron-density map (Fig. 1a–c) that reveals the detailed structural architecture of this LH1–RC complex and the relative locations of all pigments, cofactors and subunits. The dimensions of the LH1–RC are shown in Fig. 1c, d. The height of the core complex from the top of the periplasmic cytochrome to the bottom of the H subunit on the cytoplasmic side is 128.9 Å (Fig. 1a, d), and the diameters of this structure, which is slightly elliptical in projection, are 120.2 and 124.5 Å (Fig. 1c). The complex has a molecular weight of 414 kDa.

The RC in the cryo-EM map is similar to the one in the X-ray structure (for example, PDB: 1PRC)[21]. The RC consists of H, M, L and cytochrome (C) subunits. Structural differences, indicated by

---

[1]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK. [2]Electron Bio-imaging Centre, Diamond Light Source, Didcot, UK. [3]Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, UK. *e-mail: p.qian@sheffield.ac.uk; c.n.hunter@sheffield.ac.uk
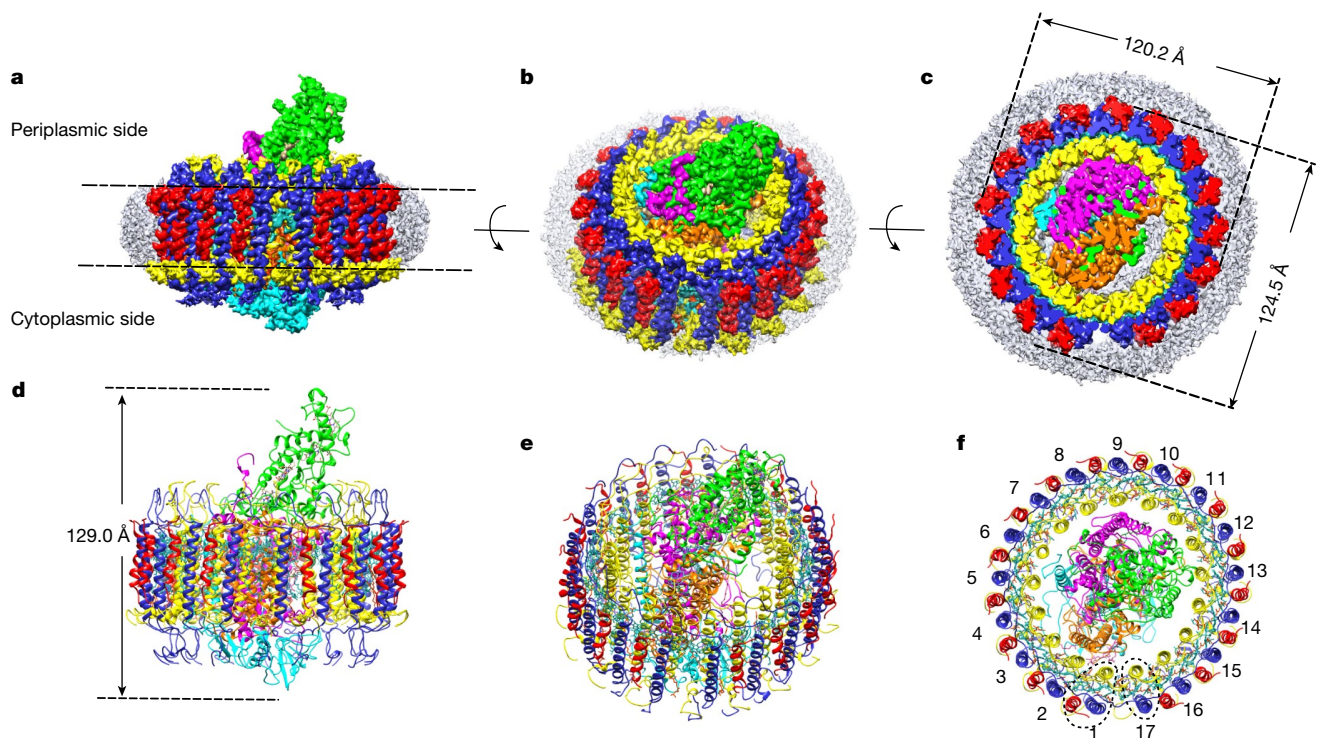
Fig. 1 | Cryo-EM structure of the LH1–RC core complex from *Blc. viridis*. a–c, Views of the colour-coded LH1–RC density map. LH1-α (yellow), LH1-β (dark blue), LH1-γ (red), BChl *b* (light sea green), carotenoid (orange red), RC-C (green), RC-H (cyan), RC-L (orange) and RC-M (magenta). Detergent and other disordered molecules are in grey. **a,** View in the plane of the membrane; two dashed lines indicate the likely position of the membrane bilayer. **b,** Forty-five-degree rotation of **a**. **c,** Perpendicular view from the periplasmic side. Densities outside the membrane region have been truncated for clarity. **d–f,** Ribbon models corresponding to **a**, **b** and **c** but without the truncations in **f**. The LH1 subunits are numbered in **f**. Subunits 1 and 17 are outlined with dashed lines in **f**.

residue–residue distance deviation[22], are small in subunits C, M and L (Extended Data Fig. 2b–d). However, interaction with the LH1 complex constrains a loop region on RC subunit H (RC-H; residues 47–54), resulting in a larger deviation from the RC-only structure (Extended Data Fig. 2a, e). There is also a small displacement of subunits RC-C and RC-H, which is likely to be caused by interaction with the LH1 complex, which bends the RC via a hinge point near the interface between the RC-C and RC-M–RC-L subunits (Extended Data Fig. 2a). The LH1 complex encircles the RC, including subunits C, H, M and L, the structures of which are in agreement with previous studies[23] (Fig. 1b, c, e, Extended Data Fig. 2).

The LH1 complex surrounds the RC to form a closed elliptical ring. The lengths of the major and minor axes of the elliptical ring measured from centre to centre of the transmembrane helices are 75.2 and 78.7 Å for the α-ring, 107.5 and 111.7 Å for the β-ring, and 109.6 and 114.8 Å for the γ-ring, respectively. The LH1 ring consists of 17 components, with 16 heterotrimers of α–β–γ-polypeptides and one α–β-heterodimer (Fig. 1c, f). Each α-, β- and γ-polypeptide contains a single transmembrane helix. A short N-terminal helix in α-polypeptides runs parallel to the membrane surface, whereas the C-terminal region contains a loop structure. No helical structures are observed in the C- and N-terminal regions of the β-polypeptide. The N termini of the α- and β-polypeptides are on the cytoplasmic side of the membrane, but the γ-subunit has the opposite topology, with its N terminus on the periplasmic side (Extended Data Fig. 3). This arrangement of LH1 polypeptides creates a triple-ring LH1 complex, consisting of an inner circle of 17 α-polypeptides, an outer ring of 16 γ-polypeptides, and a 17 β-polypeptide ring sandwiched between them (Fig. 1c, f). Each of the 16 γ-polypeptides sits between two β-polypeptides, and the gap where the 'missing' 17th γ-polypeptide would otherwise be located creates an opening in the LH1 ring (Fig. 1c, f) for quinol exchange.

Two BChl *b* molecules and one carotenoid, an all-*trans*-1,2-dihydro-derivative of neurosporene ($n = 9$) or lycopene ($n = 11$), are non-covalently bound between each α–β-pair, and no pigment molecules are bound to the γ-polypeptide (Fig. 3). Major cofactors bound within the RC are as previously reported[21] with the addition of the ubiquinone-9 Q_P (Fig. 2), and are arranged in the expected local pseudo-two-fold rotational symmetry (Fig. 2).

Interactions that stabilize the LH1 ring

The cryo-EM model of the LH1–RC from *Blc. viridis* reveals a complicated interconnecting series of protein–protein, pigment–protein and pigment–pigment associations within the LH1 ring. For simplicity, the LH1 heterotrimer subunits 1, 2 and 3 are used to demonstrate the stabilizing intra- and inter-subunit interactions in the LH1 complex. Inter-subunit hydrogen bonds on the periplasmic side are $\alpha(n)$-Arg44 to $\beta(n-1)$-Val55 (bond length 3.0 Å); and $\beta(n)$-Arg44 to $\beta(n-1)$-Ala48 (3.3 Å) (Fig. 3a). There is an intra-subunit hydrogen bond between α-Arg44 on the periplasmic side and the carboxyl group of β-Trp46 (3.0 Å), which stabilizes the C-terminal loops of both the α- and β-polypeptides (Fig. 3b). The $\gamma(n)$-polypeptide forms two hydrogen bonds with the $\alpha(n)$- and $\beta(n)$-polypeptides: γ-Asp14 to β-Trp41 (3.0 Å) and γ-Arg36 to the carboxyl group of α-Thr6 (3.1 Å) (Fig. 3b). Thus, an LH1 heterotrimer subunit is formed from $\alpha(n)$–$\beta(n)$–$\gamma(n)$, and not $\alpha(n+1)$–$\beta(n+1)$–$\gamma(n)$. This arrangement suggests an assembly sequence for the LH1 complex. It is likely that once an $\alpha(1)$–$\beta(1)$ subunit is formed, it interacts with RC-H to form an anchor point through the hydrogen bond between $\alpha(1)$-Arg19 and RC-H-Ser256. Then, the $\gamma(1)$-polypeptide binds to the α–β-subunit to form the first LH1 subunit $\alpha(1)$–$\beta(1)$–$\gamma(1)$. To do so, the γ-polypeptide needs space to access the α–β-subunit by rotating and translating to achieve the correct angle of approach and the correct orientation. This procedure continues until the 17th α–β-subunit is assembled. At this point, there is no space for a correct direction of approach and orientation that
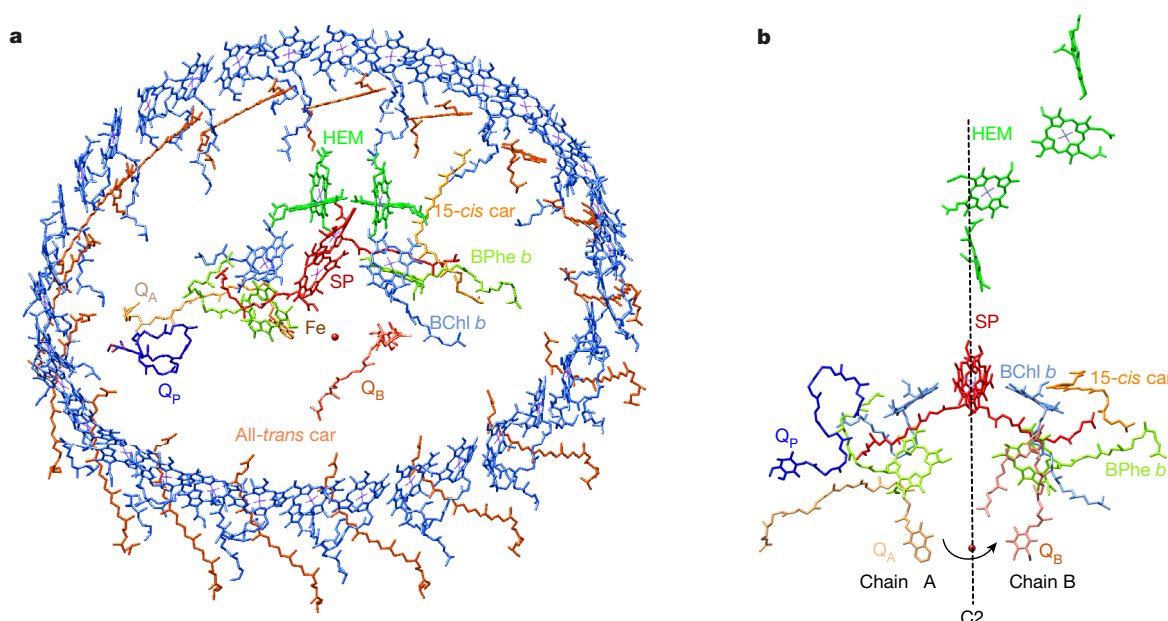
Fig. 2 | Pigment arrangement in the *Blc. viridis* LH1–RC core complex. **a**, Pigment molecules viewed from the periplasmic side by tilting 45° in the plane of the membrane. **b**, RC pigment molecules viewed from the membrane plane. A local pseudo-C2 symmetry axis is shown as a dashed line. Car, carotenoid; HEM, haem cofactors of the cytochrome subunit; SP, special pair of BChl *b* pigments; Fe, non-haem iron; BPhe *b*, bacteriopheophytin *b*.

would allow the 17th γ-polypeptide to dock with the 17th α–β-subunit, resulting in a gap in the LH1 ring.

The LH1–RC from *Blc. viridis* reveals the basis for the stabilizing effects of carotenoids, which mainly rely on hydrophobic forces, and for excitation energy transfer from carotenoids to BChls[15]. Interactions of each carotenoid with $n+1$, n and $n-1$ polypeptides and with bound BChls effectively crosslink one LH1 α–β-subunit to the next (Fig. 3c). One end of the carotenoid is in close proximity to the upstream neighbouring LH1 α($n+1$) near its C terminus (Phe37, 3.1 Å; Leu33, 3.7 Å; Ala32, 3.4 Å; His36, 3.9 Å); the other end approaches the downstream neighbouring LH1 α($n-1$) near its N terminus (Leu11, 4.3 Å; Lys10, 5.1 Å). In particular, this end of the

carotenoid is also in close proximity to the β(n)-N terminus. The middle part of the carotenoid is close to the phytyl tails of the α- (3.2 Å) and β- (4.0 Å) BChl *b* molecules (Fig. 3c).

Subunits 1–16 of the LH1 complex (Fig. 1f) consist of one each of α-, β- and γ-polypeptides, two BChl *b* molecules and one all-*trans* carotenoid. The γ-polypeptide has no histidine residue and does not bind BChl *b*. Figure 3b illustrates this point using subunit 3; α-His36 forms a ligand with the central Mg of α-BChl *b* (2.5 Å) and β-His37 forms a ligand with β-BChl *b* (2.2 Å) (Fig. 3b, c). The C3-acetyl groups of α-BChl *b* and β-BChl *b* form a hydrogen bond with α-Trp47 (2.9 Å) and β-Trp46 (2.9 Å), respectively, to orientate the bacteriochlorin rings of BChl *b*. This orientation is further stabilized by a 3.0 Å hydrogen
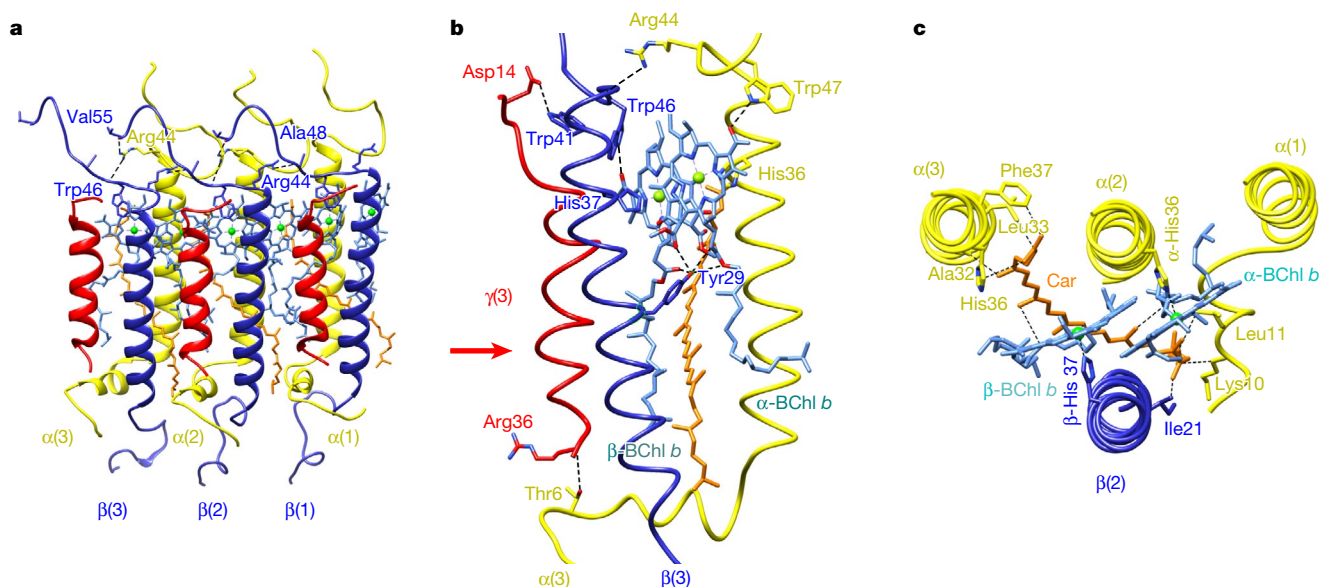

Fig. 3 | Intra- and inter-subunit protein–protein and protein–pigment interactions. **a**, LH1 subunits 1–3 (Fig. 1f) illustrate inter-subunit interactions. Colours as in Fig. 1 except BChl *b* molecules in medium blue and all-*trans*-1,2-dihydroneurosporene in orange. Hydrogen bonds are indicated by dashed lines. **b**, A single LH1 α–β–γ-subunit, with the polypeptides shown in loop representation for clarity. The red arrow indicates a putative direction of approach of the γ-polypeptide to the α–β-pair during assembly of the complex. **c**, Projection view to show interactions made by a carotenoid with nearby pigments and polypeptides.
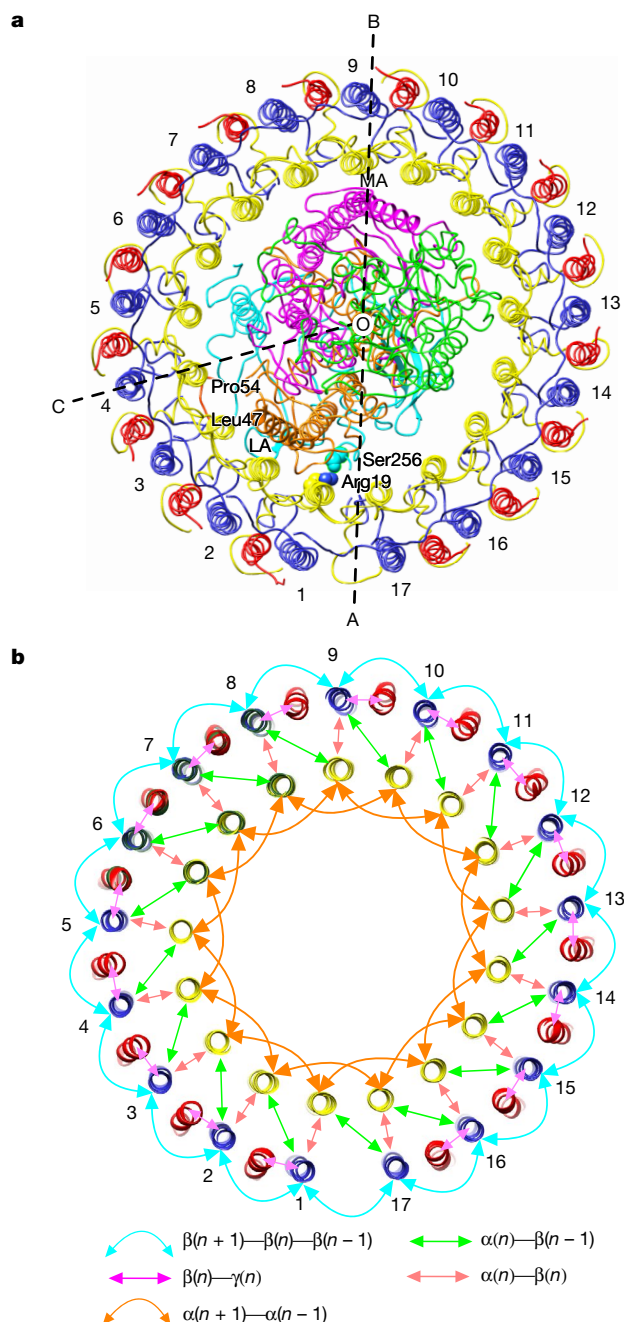
Fig. 4 | Interactions between the RC and the LH1 complex, and within the LH1 complex. a, Periplasmic side of the LH1–RC complex; colour coding as in Fig. 1. RC-H Ser256 and LH1-α1 Arg19 are highlighted using space-fill. The RC-H-loop Leu47 to Pro54 is highlighted in orange. **b,** Summary of intra- and inter-subunit interactions in the LH1 complex. Only transmembrane helices of LH1 polypeptides are shown for clarity. All arrows indicate hydrogen-bonding interactions.

bond between β-Tyr29 and the ester group of β-BChl b on C13^2. The OH group of β-Tyr29 could form a hydrogen bond with the ester group on the phytyl tail of the α- or β-BChl b.

LH1–RC interactions

The resolution of the cryo-EM structure of the LH1–RC complex is sufficient to enable detailed analysis of the protein–protein and protein–pigment interactions within the complex. The protein–pigment interactions in the RC have previously been described in detail[24], and the relationship between the RC and its encircling LH1 can now be defined. Figure 4a shows the overall organization of the LH1–RC complex, which is divided into three zones. Zone 1 (AOC) includes

a close contact between LH1 and the RC, a hydrogen bond between LH1-α1-Arg19 and RC-H-Ser256 (2.8 Å), which is likely to be the site for initiating encirclement by LH1 in a manner analogous to that in the LH1–RC–PufX complex of *Rba. sphaeroides*[6]. This trimeric α–β–γ subunit is designated as LH1 subunit 1 (Fig. 1f). Proximity between transmembrane helix RC-L$_A$ and LH1-α2, with a centre–centre helix distance of approximately 10 Å, could facilitate the encirclement process. A third interaction in this region involves LH1-α3 and LH1-α4 on the cytoplasmic side, which constrains a loop on RC-H (Leu47–Pro54). In zone 2 (COB) there is a single point of contact between the RC and LH1, between the RC-M$_A$ helix and LH1-α9. The gap between the RC and LH1 in this region is mainly filled by the single transmembrane helix of RC-H and lipid molecules (Extended Fig. 4a). Zone 3 (BOA) is where quinol–quinone exchange occurs at the RC Q$_B$ site, and where newly released quinols and quinones arriving from outside create a dynamic quinone pool[6]. Thus, the structure of the gap between the RC and LH1 in this region shows disordered densities arising from lipids and quinones[5] (Extended Data Fig. 4a). Figure 4b summarizes all intra- and inter-subunit protein–protein and protein–pigment interactions in the LH1 complex, and highlights the extent of the interactions that stabilize the LH1 complex.

Structural basis for the redshift

The LH1–RC complex of *Blc. viridis* is able to absorb energy in the infrared region of the spectrum owing to the unusually large redshift it imposes on the BChl b pigment; its 1,015-nm absorption maximum represents, to our knowledge, the lowest energy light used by a photosynthetic bacterium. Previous studies have identified several influences on the redshift of the BChl a or BChl b–Q$_y$ absorption maximum in the bacterial light-harvesting complex[25, 26]. The cryo-EM structure of the *Blc. viridis* LH1–RC complex shows that at least five factors contribute to the large bathochromic shift of the BChl b-Q$_y$ band.

The first factor is the chemical structure. The extra C–C double bond in BChl b relative to BChl a extends conjugation in the bacteriochlorin ring and redshifts the Q$_y$ band. The 795-nm absorption maximum of BChl b in methanol is 24 nm further towards the near infrared than that of BChl a, which directly affects the 'site energy' within coupled BChl b aggregates in the LH1–RC complex.

The second factor is protein–pigment interactions. As already noted (Fig. 3c), the carotenoids interlink LH1 α–β-subunits (Fig. 3c) and the C3 acetyl groups of α- and β-BChls b hydrogen bond to Trp residues in LH1 (Fig. 3b), adopting an in-plane conformation similar those of the B800-850 BChls a in the LH2 complex of *Rhodopseudomonas acidophila*[27, 28]. Experiments combining mutagenesis and Raman spectroscopy have shown that hydrogen bonds redshift the absorption of the *Rba. sphaeroides* LH1 complex[29, 30].

The third factor is the number of coupled BChl a and b molecules; the 17 pairs of coupled BChl b molecules in the LH1–RC complex of *Blc. viridis* represent the largest reported circular aggregate of pigments in light-harvesting complexes from photosynthetic bacteria[31]. Increasing the oligomeric size of LH1 subunits in the LH1 complex of *Rba. sphaeroides* from 2 to 6 or 7 is accompanied by redshifts of 6–7 nm in absorption and fluorescence emission of the BChl a-Q$_y$ band, although larger oligomers produced no further redshifts[32].

The fourth factor is the structures of BChl a or BChl b aggregates. The Mg–Mg distances within BChl pairs reflect the degree of overlap, and therefore the electronic coupling and Q$_y$ redshifting, of BChl a and BChl b in light-harvesting complexes. Extended Data Fig. 5 shows the linear correlation of Q$_y$-band maximum versus inter- and intra-subunit Mg–Mg distances in five different light-harvesting complexes, which shows a stronger correlation for the intra-subunit distances. The intra-subunit (8.8 Å) or inter-subunit (8.5 Å) Mg–Mg distances of BChl b in *Blc. viridis* are the shortest reported for a bacterial light-harvesting complex.

Finally, the shift is affected by the structural rigidity enforced by the γ-apoproteins. Sixteen γ-apoproteins pack tightly between β-polypeptides, and collectively interlock the LH1 structure through
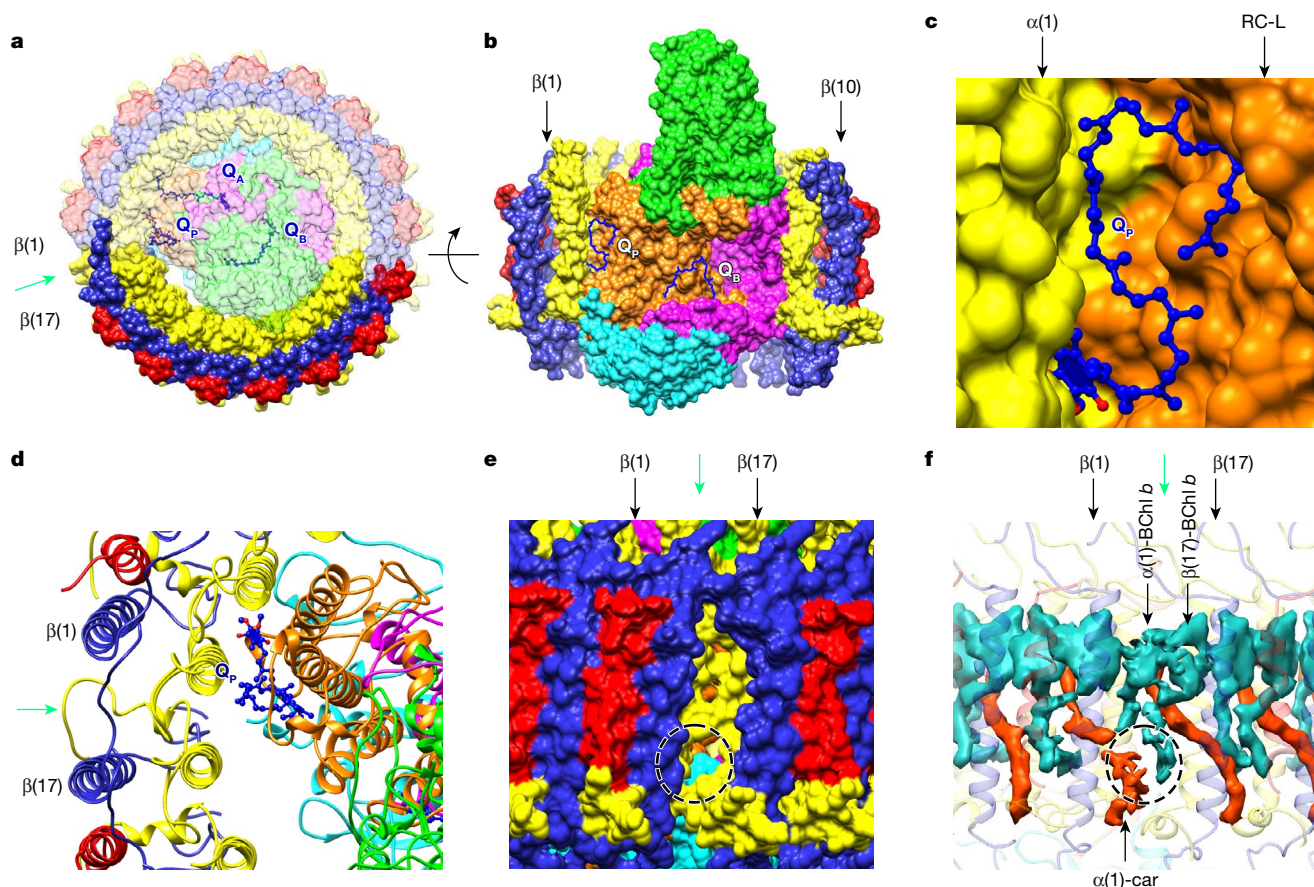
Fig. 5 | A quinone–quinol channel in the LH1–RC core complex.
a, The LH1–RC complex viewed from the periplasmic side, with 80%
transparency applied to the RC, and LH1 subunits 9–17. A green arrow
indicates the gap between subunits 1 and 17. **b**, LH1–RC rotated 90° from
a, with Q_B and Q_P viewed by removing LH1 subunits 1–8. **c**, Close-up of
the Q_P-binding pocket. **d**, Ribbon representation of the Q_P region. Green
arrow as in **a**. **e**, Close-up view of the quinone–quinol channel (dashed
circle) from outside the LH1 ring. **f**, Electron densities of pigments
adjacent to the LH1 pore.

32 hydrogen bonds to α- and β-polypeptides, constraining free move-
ment of the LH1 ring and stabilizing the BChl *b* pairs in the complex
and thereby contributing to the redshift of the BChl *b*-Q_y band[33].
There are parallels with the large redshift of BChl *a* to 915 nm in
the LH1–RC complex from *Tch. tepidum* (Extended Data Fig. 6a).
In this case, bound Ca^{2+} ions constrain conformational flexibility[34]
and limit disorder in site energies. Inhomogeneous narrowing is
accompanied by mixing of charge transfer and lowest exciton states,
which has been hypothesized to be the basis for the redshift in this
complex[25].

A template for preparing quinols for export
The LH1–RC of *Blc. viridis* houses the RC quinones Q_A and Q_B, and
a third quinone, Q_P (Fig. 5a–c). The binding sites of Q_A and Q_B are
similar to those reported previously, although their tail structures differ
from those in isolated RCs[21,35] (Extended Data Fig. 6b, c). Q_P is located
near the gap in the LH1 ring, 48.9 Å away from the Q_B-binding site. The
head of the Q_P molecule is stabilized by π–π-interactions with RC-L-
Phe40, the aromatic ring of which is roughly parallel to the plane of the
quinone-head group at a distance of 3.6 Å. Q_P is also in close proximity
(3.0 Å) to LH1-α1-Tyr27, the aromatic ring of which is roughly per-
pendicular to the plane of the Q_P-head ring. Unlike Q_A and Q_B in the
RC, the tail of Q_P is not free to move, and is instead conformationally
constrained by a series of contacts with LH1-α(1)-Phe37 (4.7 Å), RC-L-
Gln87 (2.4 Å), RC-L-Trp142 (3.5 Å) and RC-L-Val91(4.4 Å) (Extended
Data Fig. 6d). This Q_P-binding pocket provides a folding template such
that Q_P assumes a compact conformation and a suitable orientation
before entering the pore in LH1 at the position of the absent 17th
γ-apoprotein (Fig. 5c, d).

The cryo-EM structure of the LH1–RC complex reveals the mech-
anism by which quinone is translocated across the LH1 ring. Of the
17 subunits, 16 are α–β–γ-heterotrimers, and only one is an α–β-
heterodimer. The 16 γ-polypeptides, located outside the β-ring, pack
between β-apoproteins, leaving a gap in the LH1 ring between subu-
nits 1 and 17, and dictate the position of the pore for quinone–quinol
translocation. The Q_P-binding pocket is located next to the pore, and
the Q_P molecule appears to be folded and oriented in the binding
pocket in a manner that encourages passage through the LH1 ring
(Fig. 5d). There is a distinct pore measuring around 5×7 Å between
α17 and α1 (Fig. 5e), which is created by Arg18–Phe25 (sequence,
RRVLTALF) in α17 and Leu15–Leu24 (LDPRRVLTAL) in α1
(Fig. 5e). It should be noted that the electron densities of β(17)-BChl
b, α(1)-BChl *b* and the α(1)–β(1) carotenoid are weaker than those of
their counterparts in the rest of the LH1 complex. This is particularly
evident for those regions of the pigments that are close to the quinone
pore, for example the phytyl tails and one end of the carotenoid, as
shown in Fig. 5f. This weaker density reflects the relative flexibility of
this region; thus the size of this pore could fluctuate transiently, faci-
litating the movement of the quinone and quinol molecules through
the channel.

Online content
Any Methods, including any statements of data availability and Nature Research
reporting summaries, along with any additional references and Source Data files,
are available in the online version of the paper at https://doi.org/10.1038/s41586-
018-0014-5.

1. Blankenship, R. E. *Molecular Mechanisms of Photosynthesis* 2nd edn (John Wiley & Sons, Oxford, 2014).
2. Sundström, V., Pullerits, T. & van Grondelle, R. Photosynthetic light-harvesting: Reconciling dynamics and structure of purple bacterial LH2 reveals function of photosynthetic unit. *J. Phys. Chem. B* **103**, 2327–2346 (1999).
3. Cogdell, R. J., Gall, A. & Köhler, J. The architecture and function of the light-harvesting apparatus of purple bacteria: from single molecules to in vivo membranes. *Q. Rev. Biophys.* **39**, 227–324 (2006).
4. Niwa, S. et al. Structure of the LH1–RC complex from *Thermochromatium tepidum* at 3.0 Å. *Nature* **508**, 228–232 (2014).
5. Jamieson, S. J. et al. Projection structure of the photosynthetic reaction centre-antenna complex of *Rhodospirillum rubrum* at 8.5 A resolution. *EMBO J.* **21**, 3927–3935 (2002).
6. Roszak, A. W. et al. Crystal structure of the RC–LH1 core complex from *Rhodopseudomonas palustris*. *Science* **302**, 1969–1972 (2003).
7. Jungas, C., Ranck, J. L., Rigaud, J. L., Joliot, P. & Verméglio, A. Supramolecular organization of the photosynthetic apparatus of *Rhodobacter sphaeroides*. *EMBO J.* **18**, 534–542 (1999).
8. Qian, P., Hunter, C. N. & Bullough, P. A. The 8.5 Å projection structure of the core RC–LH1–PufX dimer of *Rhodobacter sphaeroides*. *J. Mol. Biol.* **349**, 948–960 (2005).
9. Qian, P. et al. Three-dimensional structure of the *Rhodobacter sphaeroides* RC–LH1–PufX complex: dimerization and quinone channels promoted by PufX. *Biochemistry* **52**, 7575–7585 (2013).
10. Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. X-ray structure analysis of a membrane protein complex. Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. *J. Mol. Biol.* **180**, 385–398 (1984).
11. Miller, K. R. & Jacob, J. S. The *Rhodopseudomonas viridis* photosynthetic membrane: arrangement in situ. *Arch. Microbiol.* **142**, 333–339 (1985).
12. Ikeda-Yamasaki, I., Odahara, T., Mitsuoka, K., Fujiyoshi, Y. & Murata, K. Projection map of the reaction center-light harvesting 1 complex from *Rhodopseudomonas viridis* at 10 Å resolution. *FEBS Lett.* **425**, 505–508 (1998).
13. Blankenship, R. E. et al. Comparing photosynthetic and photovoltaic efficiencies and recognizing the potential for improvement. *Science* **332**, 805–809 (2011).
14. Magdaong, N. C. M., Niedzwiedzki, D. M., Goodson, C. & Blankenship, R. E. Carotenoid-to-bacteriochlorophyll energy transfer in the LH1–RC core complex of a bacteriochlorophyll *b* containing purple photosynthetic bacterium *Blastochloris viridis*. *J. Phys. Chem. B* **120**, 5159–5171 (2016).
15. Thornber, J. P., Cogdell, R. J., Seftor, R. E. B. & Webster, G. D. Further studies on the composition and spectral properties of the photochemical reaction centers of bacteriochlorophyll *b*-containing bacteria. *Biochim. Biophys. Acta* **593**, 60–75 (1980).
16. Malhotra, H. C., Britton, G. & Goodwin, T. W. Occurrence of 1,2-dihydro-carotenoids in *Rhodopseudomonas viridis*. *J. Chem. Soc. Chem. Comm.* **127**, 127–128 (1970).
17. Konorty, M. et al. Photosynthetic system in *Blastochloris viridis* revisited. *Biochemistry* **48**, 4753–4761 (2009).
18. Welte, W. & Kreutz, W. Structure of thylakoids in cells of *Rhodopseudomonas viridis* as influenced by growth conditions. *Arch. Microbiol.* **141**, 325–329 (1985).
19. Engelhardt, H., Baumeister, W. & Saxton, W. O. Electron microscopy of photosynthetic membranes containing bacteriochlorophyll *b*. *Arch. Microbiol.* **135**, 169–175 (1983).
20. Scheuring, S. et al. Nanodissection and high-resolution imaging of the *Rhodopseudomonas viridis* photosynthetic core complex in native membranes by AFM. *Proc. Natl Acad. Sci. USA* **100**, 1690–1693 (2003).
21. Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J. Mol. Biol.* **246**, 429–457 (1995).
22. Chen, J. E., Huang, C. C. & Ferrin, T. E. RRDistMaps: a UCSF Chimera tool for viewing and comparing protein distance maps. *Bioinformatics* **31**, 1484–1486 (2015).
23. Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature* **318**, 618–624 (1985).
24. Deisenhofer, J. & Michel, H. Nobel lecture. The photosynthetic reaction centre from the purple bacterium *Rhodopseudomonas viridis*. *EMBO J.* **8**, 2149–2170 (1989).
25. Ma, F., Yu, L. J., Wang-Otomo, Z. Y. & van Grondelle, R. The origin of the unusual Q_y red shift in LH1–RC complexes from purple bacteria *Thermochromatium tepidum* as revealed by Stark absorption spectroscopy. *Biochim. Biophys. Acta* **1847**, 1479–1486 (2015).
26. Cogdell, R. J., Howard, T. D., Isaacs, N. W., McLuskey, K. & Gardiner, A. T. Structural factors which control the position of the Q_y absorption band of bacteriochlorophyll *a* in purple bacterial antenna complexes. *Photosynth. Res.* **74**, 135–141 (2002).
27. McDermott, G. et al. Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **374**, 517–521 (1995).
28. McLuskey, K., Prince, S. M., Cogdell, R. J. & Isaacs, N. W. The crystallographic structure of the B800–820 LH3 light-harvesting complex from the purple bacteria *Rhodopseudomonas acidophila* strain 7050. *Biochemistry* **40**, 8783–8789 (2001).
29. Olsen, J. D., Sockalingum, G. D., Robert, B. & Hunter, C. N. Modification of a hydrogen bond to a bacteriochlorophyll *a* molecule in the light-harvesting 1 antenna of *Rhodobacter sphaeroides*. *Proc. Natl Acad. Sci. USA* **91**, 7124–7128 (1994).
30. Sturgis, J. N., Olsen, J. D., Robert, B. & Hunter, C. N. Functions of conserved tryptophan residues of the core light-harvesting complex of *Rhodobacter sphaeroides*. *Biochemistry* **36**, 2772–2778 (1997).
31. Cogdell, R. J. & Roszak, A. W. Structural biology: the purple heart of photosynthesis. *Nature* **508**, 196–197 (2014).
32. Westerhuis, W. H. J., Sturgis, J. N., Ratcliffe, E. C., Hunter, C. N. & Niederman, R. A. Isolation, size estimates, and spectral heterogeneity of an oligomeric series of light-harvesting 1 complexes from *Rhodobacter sphaeroides*. *Biochemistry* **41**, 8698–8707 (2002).
33. Yu, L. J., Kawakami, T., Kimura, Y. & Wang-Otomo, Z. Y. Structural basis for the unusual Q_y red-shift and enhanced thermostability of the LH1 complex from *Thermochromatium tepidum*. *Biochemistry* **55**, 6495–6504 (2016).
34. Jakob-Grun, S., Radeck, J. & Braun, P. Ca^{2+}-binding reduces conformational flexibility of RC–LH1 core complex from thermophile *Thermochromatium tepidum*. *Photosynth. Res.* **111**, 139–147 (2012).
35. Roszak, A. W. et al. New insights into the structure of the reaction centre from *Blastochloris viridis*: evolution in the laboratory. *Biochem. J.* **442**, 27–37 (2012).

Author Contributions P.Q. and C.N.H. conceived the study. P.Q. and C.N.H. designed the experiments. P.Q., C.A.S., D.P.C. and P.W. performed the experiments. P.Q. analysed the results and generated structural models. P.Q. and C.N.H. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0014-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0014-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.Q. or C.N.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein purification. Wild-type *Blc. viridis* (DSM-133) was obtained from DSMZ. Photosynthetic cultures of *Blc. viridis* were grown in sodium succinate medium 27 (N medium) under illumination (100 μmol photons per m^2 s^{-1}) at 30 °C in 20-l screw-capped vessels, completely filled with N$_2$-sparged medium, as previously described[36]. Cells were collected when the culture reached an optical density of 1.6 at 680 nm by centrifugation at 3,290*g* for 30 min. Washed cells were broken by passage through a French press three times at 18,000 p.s.i. The crude cell lysate was applied to a two-step sucrose gradient (15% and 40% (w/w) in an ultracentrifugation tube). Photosynthetic membrane was collected at the interface of 15% and 40% sucrose after 5 h centrifugation at 100,000*g*. Membranes were pelleted and resuspended in working buffer (20 mM HEPES, pH 7.8). The optical density of the membrane was adjusted to ~100 at 1,015 nm. For solubilization of the core complexes, the optical density at 1,015 nm of the photosynthetic membrane was adjusted to 60, and 3% (w/w) *n*-dodecyl β-D-maltoside was added. This mixture was then stirred in the dark at 4 °C for 30 min. Unsolubilized material was removed by centrifugation for 1 h at 211,000*g*. The clarified supernatant was loaded on an ion exchange column pre-equilibrated with working buffer solution containing 0.03% *n*-dodecyl β-D-maltoside. The core complexes eluted at ~250 mM NaCl and were collected and concentrated. These were further purified using a Superdex 200 gel filtration column. The fractions with an absorption ratio of $A_{1,008\,nm}/A_{280\,nm}$ higher than 1.22 were pooled together and used for cryo-EM data collection.

Cryo-EM data collection. The protein concentration was adjusted to an optical density of 40 at 1,008 nm. Three microlitres of protein solution was applied to a glow-discharged holey carbon grid (Quantifoil grid R1.2/1.3, 300 mesh Cu). The grid was plunged into liquid ethane cooled by liquid nitrogen using a Leica EM GP vitrobot. Parameters were set as follows: blotting time 4 s, humidity 99%, sample chamber temperature 5 °C. The frozen grid was stored in liquid nitrogen before use. A second grid was prepared using a Quantifoil grid R3.5/1.0 covered by a thin carbon film (EM resolution, Inc.), with the protein diluted tenfold. Vitrification conditions were the same as for the first grid. Data were recorded at eBIC on a Titan Krios electron microscope with a Gatan 968 GIF Quantum with a K2 summit detector operating at 300 kV accelerating voltage, at nominal magnification of 130 k in counting mode. Movies were collected in super-resolution mode and Fourier-cropped to give a resulting calibrated pixel size of 1.06 Å at the specimen level. An energy-selecting slit of 20 eV was used. An exposure rate of 5 electrons per pixel per s was set and a fresh super-resolution gain reference was performed at this dose rate before data acquisition. A total dose of 45 electrons per Å^2 was used for movies of 20 frames. In total, 6,472 movies were collected with defocus values from 1.0 to 3.0 μm. Two typical cryo-EM images, which are averaged from motion-corrected movie frames, are shown in Extended Data Fig. 7a, b.

Data processing. All images that were empty, contained few particles, or were contaminated with ice were discarded. Dose-fractionated images were subjected to beam-induced motion correction using MotionCorr[37]. Images derived from the sum of all frames were used for further data processing using RELION 2.0[38–40]. CTF parameters were determined using gctf[41]. In total, 267,726 particles were picked manually. These particles were subjected to reference-free 2D classification. Those particles that categorized into poorly defined classes were rejected. This cleaning procedure using 2D classification was repeated three times, resulting in rejection of 9.45% of total particles. The resulting 2D classes were subjected to an initial 3D model calculation using EMAN2[42] for maximum-likelihood-based 3D classification. One of the four stable 3D classes accounting for 62.3% total particles was selected for high resolution refinement and 3D reconstruction without subtraction of detergent micelle from the raw micrographs. This resulted in a map at a global resolution of 3.3 Å. The density map was corrected for the modulation transfer function (MTF) of the Gatan K2 summit camera and further sharpened using the post-processing subroutine in RELION 2.0 using an estimated temperature factor and a mask was created using RELION 2.0 with a lowpass of 15 Å and a soft edge of 7 Å. The Fourier shell correlation (FSC) curve corrected for masking is shown in Extended Data Fig. 7c. The estimate of final resolution of 2.9 Å for the LH1–RC map was based on a FSC cut off of 0.143. ResMap[43] was used for calculation of the local resolution map (Extended Data Fig. 4b, c).

Modelling and refinement. Initially, the crystal structure of the *Blc. viridis* RC (PDB: 1PRC) was fitted to the cryo-EM map as a rigid body using the 'fit in map' routine in Chimera[44]. COOT[45] was then used for manual adjustment and real-space refinement for both polypeptides and cofactors. All amino acid sequences of polypeptides in the RC are listed in Extended Data Fig. 8. Ubiquinone-9 molecules (Q$_B$ and Q$_P$) were also fitted to the density map independently using COOT.

For LH1, the electron density of LH1 subunit 3 was selected for modelling first. On the basis of structural similarity compared with the LH1 of *Tch. tepidum*[5] and LH2 of *Rhodospirillum molischianum*[46], the locations of His residues, which ligate BChl *b* molecules in the α–β-polypeptides (Extended Data Fig. 9), were located in the density map. The fitted RC was used as a reference to determine the orientation of the α–β-polypeptides. Their amino acid sequences, taken from previous work[47], were fitted into an electron-density map using COOT. Two BChl *b* molecules and one all-*trans* carotenoid were added into the model based on their densities. Analysis of pigment composition shows that the major carotenoid in the core complex is all-*trans*-1,2-dihydroneurosporene[15]; this carotenoid was therefore modelled into the density map. Having no His residues, the γ-polypeptide does not bind BChl *b* molecules. No 3D structural information of the γ-subunit was available, but the 2.9 Å resolution allows assignment of the larger amino acid side chains such as Trp and Tyr. By matching three Trps and one Tyr residue in the γ-polypeptide, its orientation was determined, and all other residues were traced based on the density map using COOT. Comparison with the sequence of the γ-polypeptide[47] leaves 12 N-terminal residues unaccounted for. The structure of the LH1 α–β–γ-subunit was then used as a rigid body to fit into the density map for other LH1 subunits. For the LH1 subunit 17, only α–β and pigments were used. All of the LH1 subunits then underwent real-space refinement using COOT. The final model was subjected to global refinement and minimization using REFMAC5[48]. The final refinement statistics are summarized in Extended Data Table 1. The quality of fit for the structural model within the electron-density map was validated using EMRinger[49].

Data availability. The cryo-EM density map has been deposited in the World Wide Protein Data Bank (wwPDB) under accession code EMD-3951 and the coordinates have been deposited in the Protein Data Bank (PDB) under accession number 6ET5.

36. Lang, F. S. & Oesterhelt, D. Microaerophilic growth and induction of the photosynthetic reaction center in *Rhodopseudomonas viridis*. *J. Bacteriol.* **171**, 2827–2834 (1989).
37. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
38. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
39. Scheres, S. H. W. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).
40. Scheres, S. H. W. Processing of structurally heterogeneous cryo-EM data in RELION. *Methods Enzymol.* **579**, 125–157 (2016).
41. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
42. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
43. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
44. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
45. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
46. Koepke, J., Hu, X., Muenke, C., Schulten, K. & Michel, H. The crystal structure of the light-harvesting complex II (B800-850) from *Rhodospirillum molischianum*. *Structure* **4**, 581–597 (1996).
47. Brunisholz, R. A., Jay, F., Suter, F. & Zuber, H. The light-harvesting polypeptides of *Rhodopseudomonas viridis*—the complete amino acid sequences of B1015-α, B1015-β and B1015-γ. *Biol. Chem.* **366**, 87–98 (1985).
48. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
49. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
50. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand and contrast loss in single particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
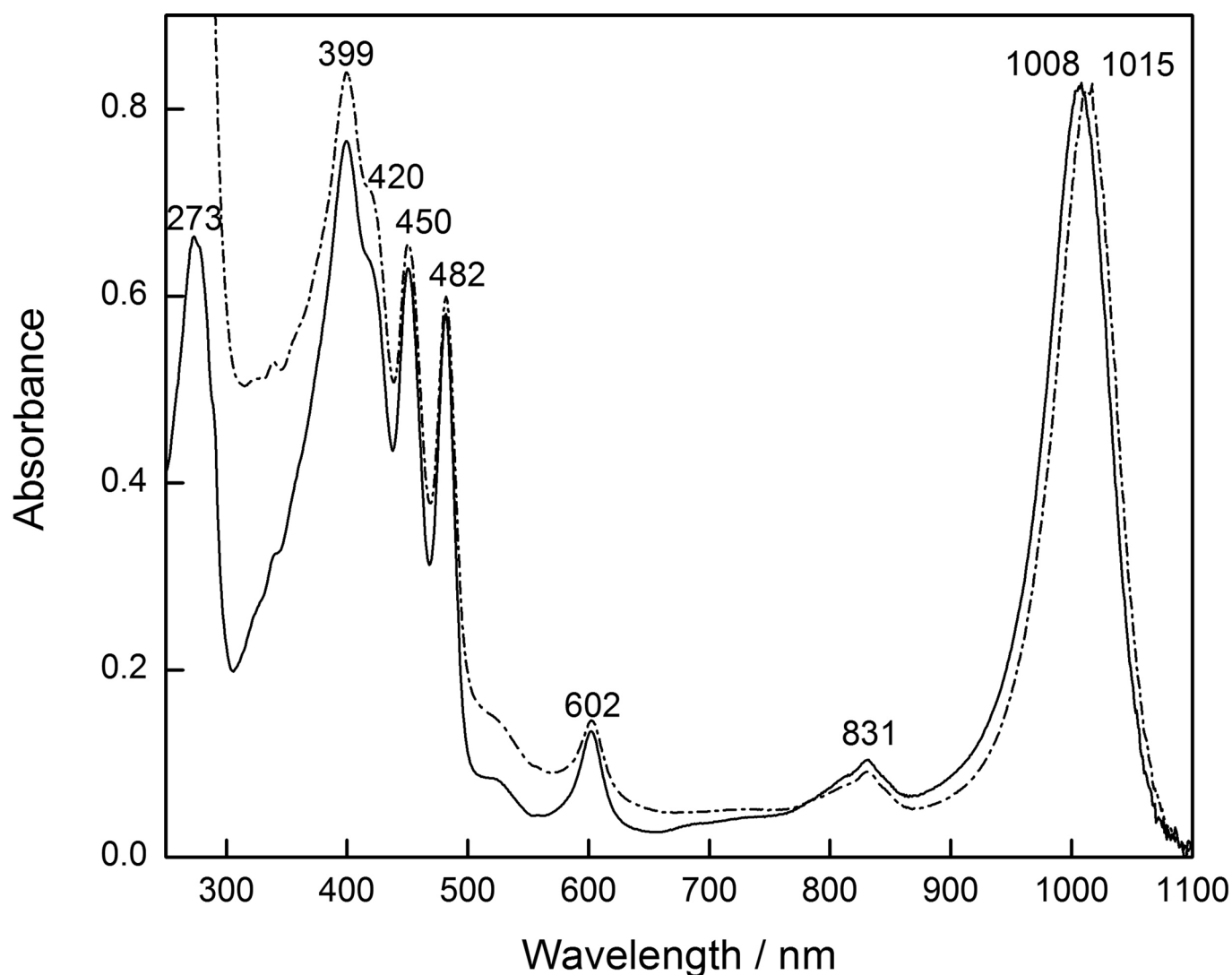
Extended Data Fig. 1 | Absorption spectra of photosynthetic membranes and the purified LH1–RC core complex from *Blc. viridis*. Absorption spectra of isolated membranes (dashed line) and the purified LH1–RC complex (solid line) were recorded at room temperature and normalized at their Q_y bands at 1,015 nm and 1,008 nm. The peak at 831 nm together with a shoulder at ~970 nm arise from BChl *b* in the RC. Bacteriopheophytin appears as a poorly resolved peak at about 810 nm. The Q_x bands give rise to a composite peak at 602 nm. The minor peak at about 558 nm arises from the cytochromes, the Soret band of which contributes in the approximately 410-nm region. Absorption features at 482, 450 and 420 nm correspond to carotenoids and the 399-nm maximum corresponds to the Soret band of BChl *b* in the core complex. No oxidized BChl *b* is observed which, if present, would cause an absorption peak at about 685 nm.
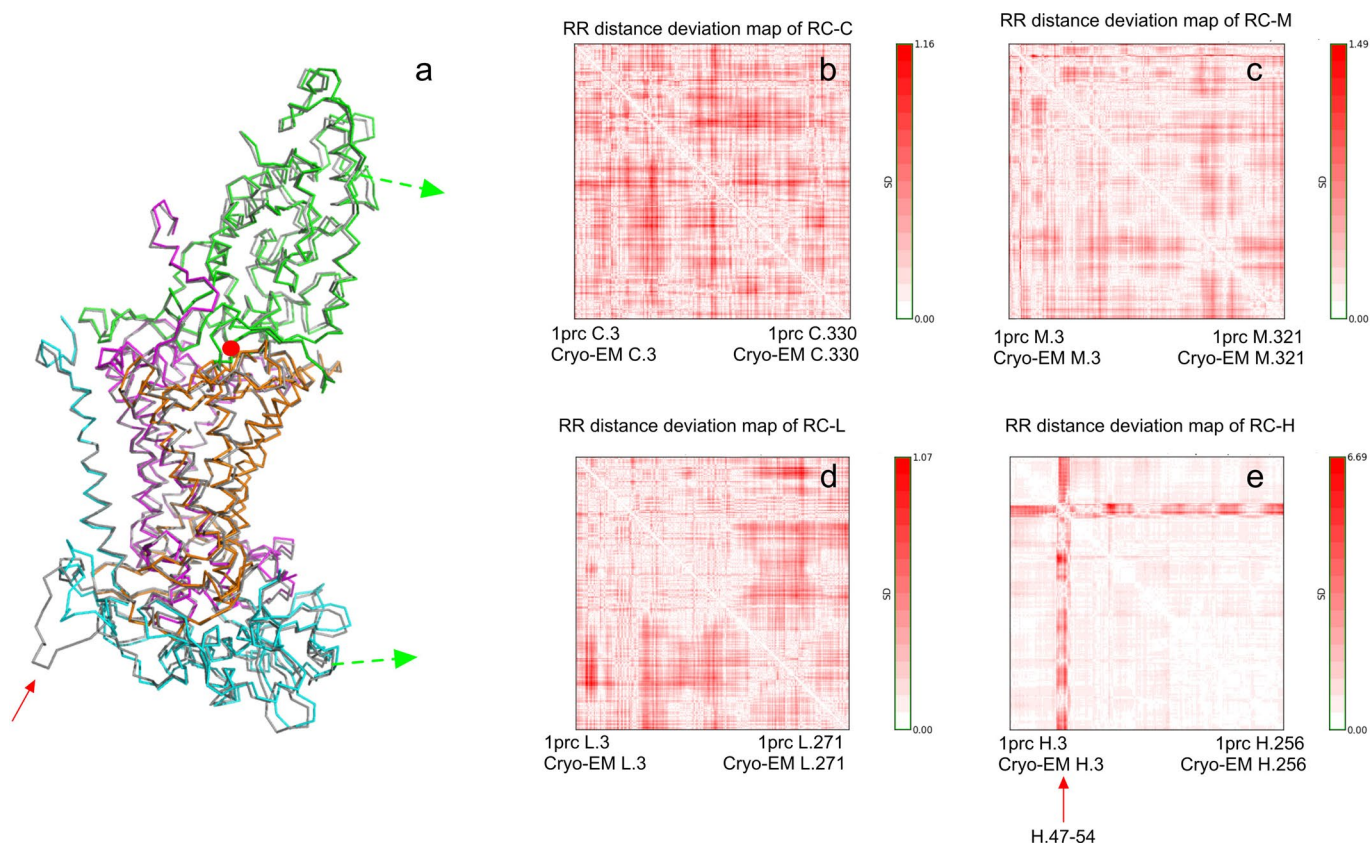
Extended Data Fig. 2 | Residue–residue distance deviation between cryo-EM and X-ray structures of the RC from *Blc. viridis*. a, Superposition of the X-ray structure (PDB: 1PRC, grey) and the cryo-EM structure (colour-coded as in Fig. 1) of the RC. A putative hinge point is indicated with a red dot. The bending direction of the cryo-EM structure is indicated with two green arrows. A red arrow points to a flexible RC-H loop. **b–e**, Residue–residue (RR) distance deviation maps[22] of the individual RC subunits C, M, L and H, respectively, comparing the structures from cryo-EM and X-ray crystallography (PDB: 1PRC)[21]. Each vertical scale shows the standard deviation (s.d.) in Å. The flexible loop of RC-H is indicated with a red perpendicular arrow in **e**.
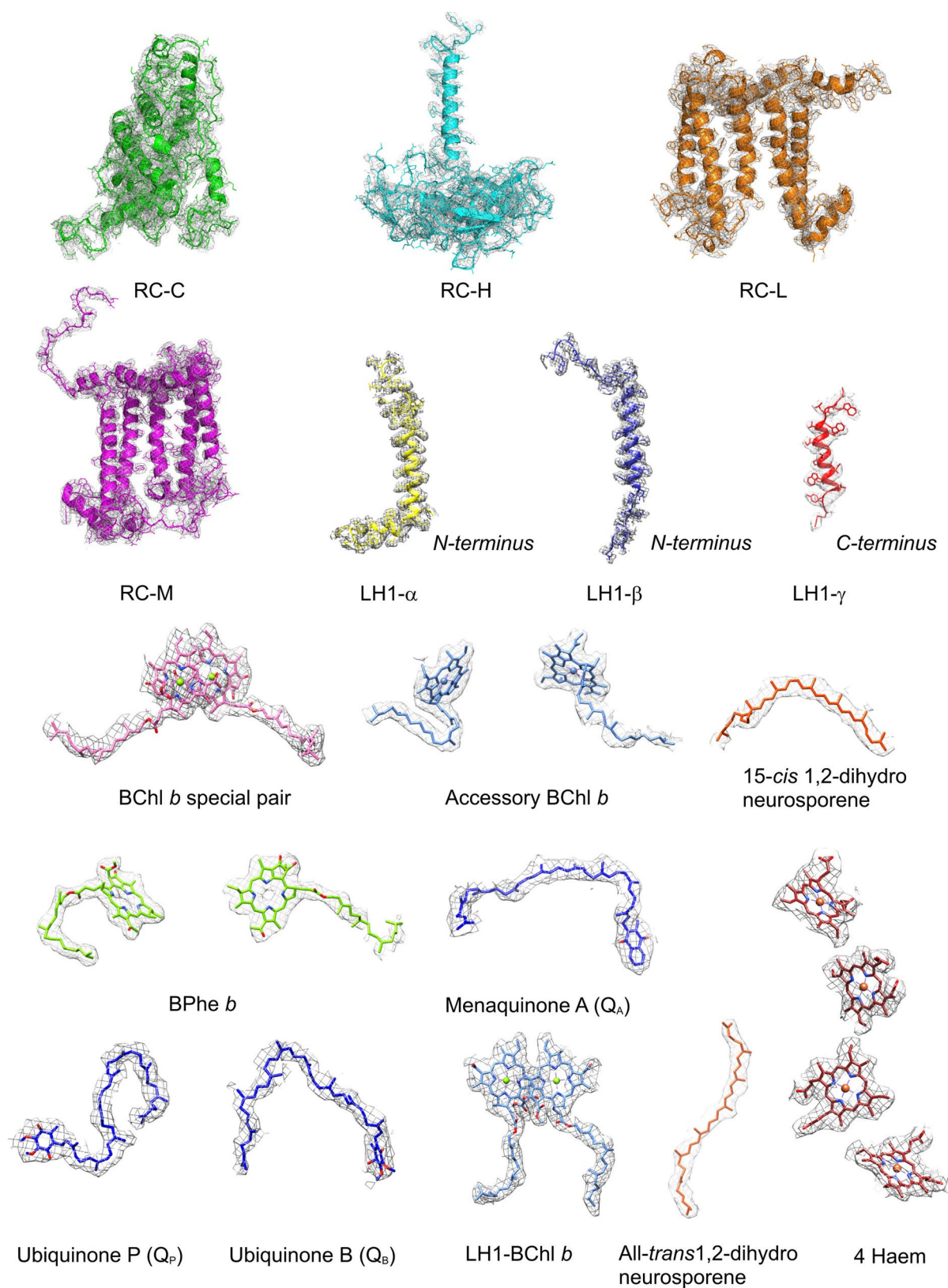
Extended Data Fig. 3 | Cryo-EM densities and structural models of polypeptides and pigments in the _Blc. viridis_ LH1–RC complex. The colour code is the same as in Fig. 1. The contour levels of the density maps were adjusted to mirror their molecular weights.
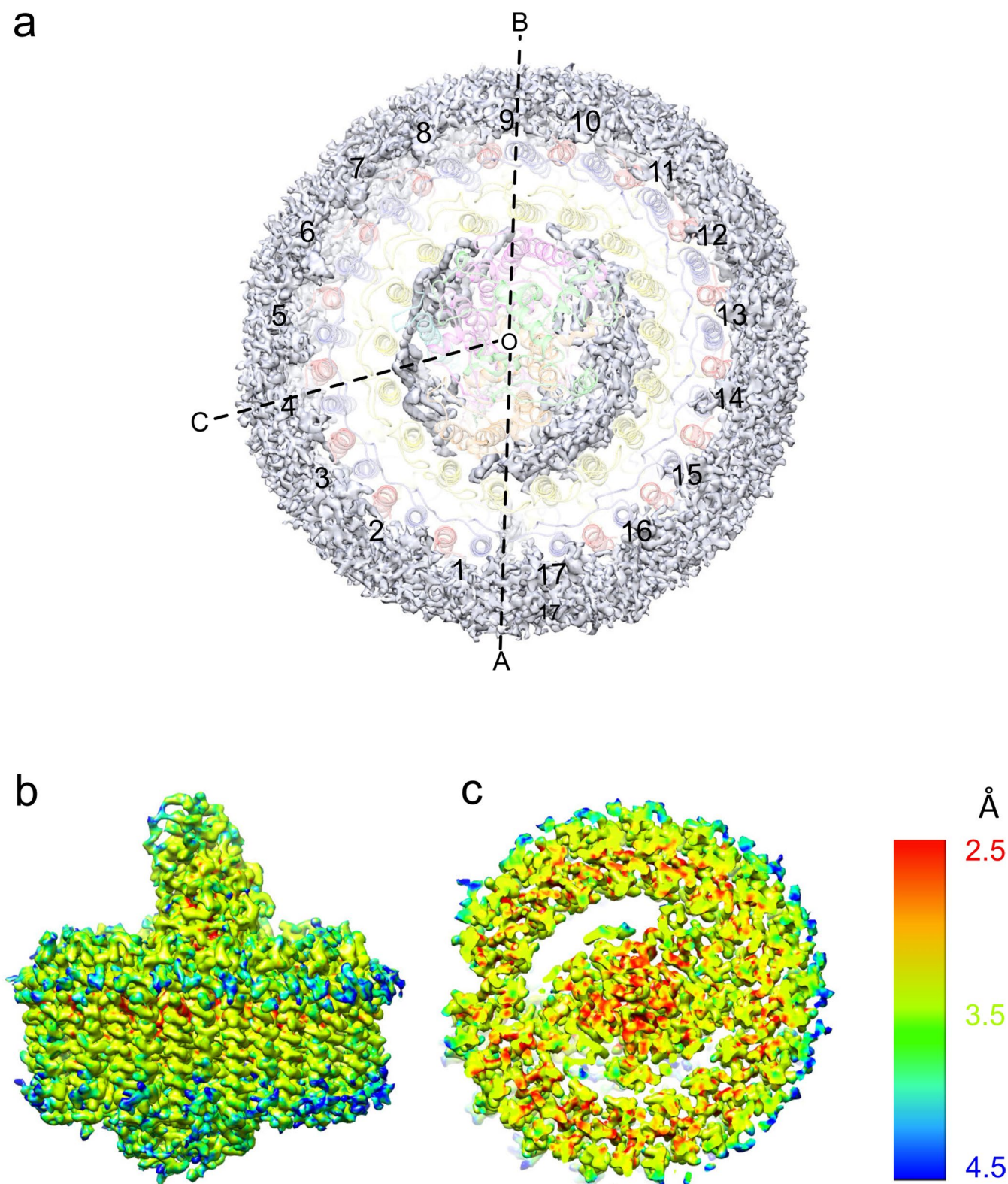
Extended Data Fig. 4 | Electron densities between and outside the LH1 and RC complexes, and local resolution maps of the LH1–RC core complex. a, The LH1–RC complex as shown in Fig. 1f, but displayed at 70% transparency. Electron densities belonging to detergent, lipid and other disordered molecules are in grey. **b**, Side view of the core complex with the periplasmic side uppermost. **c**, View of the periplasmic side. All membrane-extrinsic parts of the complex were truncated for clarity. The coloured bar chart on the right shows the local structural resolution in Å.
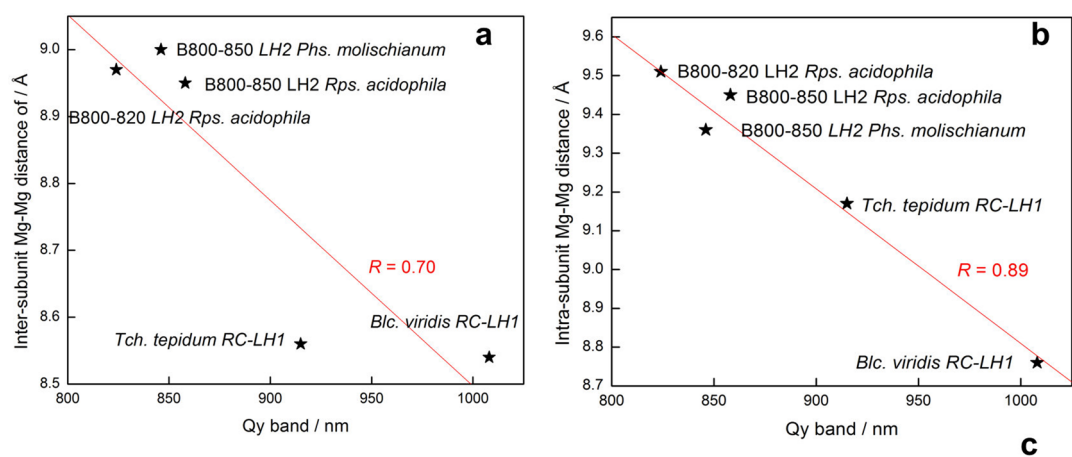
Complex	Q_Y band (nm)	Mg-Mg (Å)	
		Intra subunit*	Inter subunit*
B800-820 LH2 *Rps. acidophila*	824	9.51	8.97
B800-850 LH2 *Rps. acidophila*	858	9.45	9.00
B800-850 LH2 *Phs. molischianum*	846	9.36	8.95
RC-LH1 *Tch. tepidum*	915	8.97±0.07	8.56±0.07
RC-LH1 *Blc. viridis*	1008	8.8±0.1	8.5±0.1

*The asterisk indicates that Mg-Mg distances were calculated to two decimal places for structures obtained using X-ray crystallography. For the *Tch. tepidum* and *Blc. viridis* structures the Mg-Mg distances differ round the LH1 ring, so standard errors were calculated.

Extended Data Fig. 5 | Relationship between BChl *a* and BChl *b* Mg–Mg distances and Q$_Y$-band absorption in bacterial light harvesting complexes. a, Correlation of Q$_Y$-band maximum and inter-subunit BChl *a* and BChl *b* Mg–Mg distances in five bacterial light-harvesting complexes. **b**, As in **a**, but for intra-subunit Mg–Mg distances. **c**, Values for the linear correlation coefficient *R*, calculated using least-squares linear regression ($n = 5$ biologically independent samples in each case; one-sided significance test).
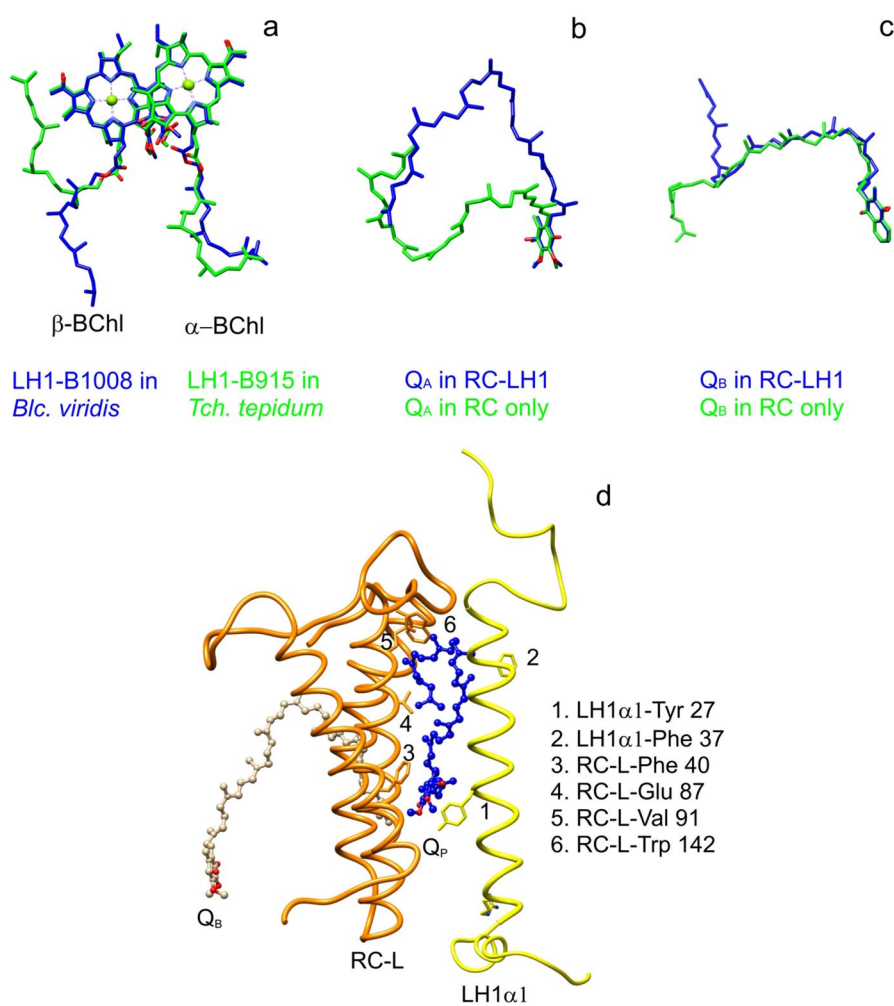
Extended Data Fig. 6 | Structural comparisons of selected cofactors and details of the Q$_P$ binding site. a, The LH1-B1008 BChl *b* pair from *Blc. viridis* (blue) compared with the LH1-B915 BChl *a* pair (green) from the X-ray structure of the *Tch. tepidum* LH1–RC complex (PDB: 3WMM). **b**, Comparison of the Q$_A$ menaquinone-9 (blue) from the cryo-EM model of the *Blc. viridis* LH1–RC with the Q$_A$ (green) from the X-ray structure of the *Blc. viridis* RC (PDB: 3T6E). **c**, As in **b**, but comparing Q$_B$. **d**, The Q$_P$ binding site. Only LH1-α1 and part of RC-L are shown for clarity. Yellow, LH1-α1; orange, RC-L; blue, Q$_P$; wheat, Q$_B$. Amino acid residues making close contacts around Q$_P$ are numbered and listed accordingly.
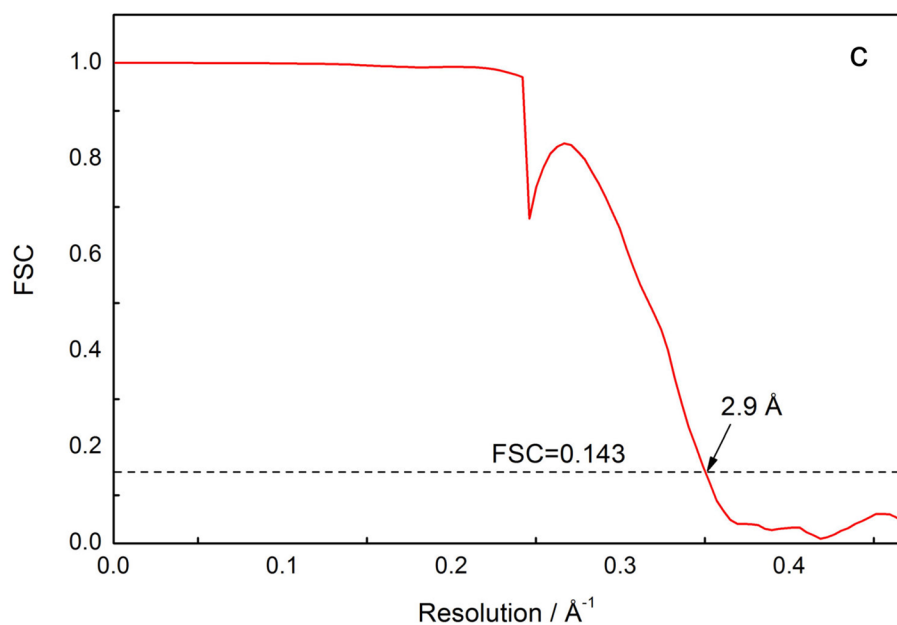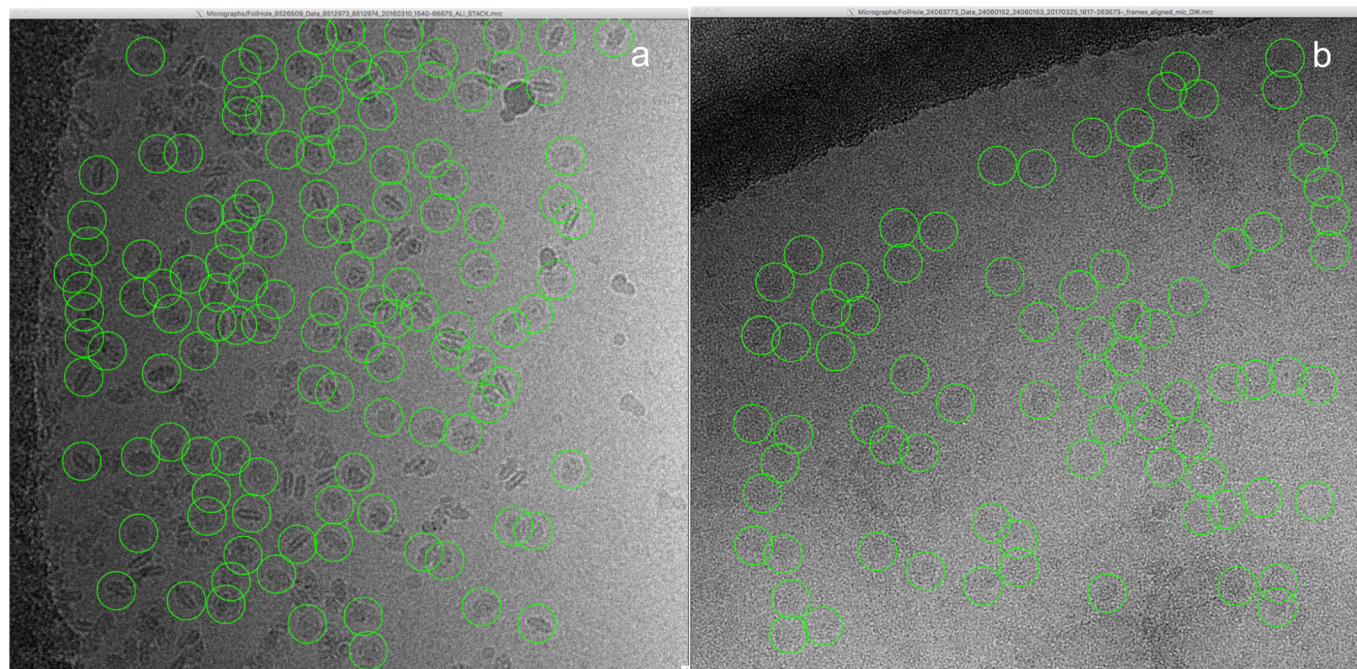
Extended Data Fig. 7 | Cryo-EM micrographs of the LH1–RC complex from *Blc. viridis* and calculation of the cryo-EM map resolution.
a, Protein particles embedded in vitrified ice. Examples of LH1–RC complexes are circled. 6,472 cryo-EM movies were recorded, from which 267,726 particles were picked manually for reference-free two-dimensional classification. During data processing, datasets of around 100,000 and around 167,000 particles were used independently for 3D reconstruction.

They generated very similar 3D maps for the LH1–RC complex, so they were combined. **b**, The LH1–RC particles are covered by a thin layer of vitrified ice on a supported carbon film. Each image measures 393.2×406.8 nm. **c**, Gold-standard refinement was used for estimation of the final map resolution. The global resolution of 2.9 Å was calculated using an FSC cut off of 0.143.

LH1 α:
MATEYRTASWKLWLILDPRRVLTALFVYLTVIALLIHFGLLSTDRLNWWEFQR GLPKAASLVVVPPAVG
1   11   21   31   41   51   61

LH1 β:
MADLKPSLTGLTEEEAKEFHGIFVTSTVLYLATAVIVHYLVWTARPWIAPIPK GWVNLEGVQSALSYLV
1   11   21   31   41   51   61

LH1 γ:
MKLSAILGALSVVLTSTIASA YFAADGSVVPSI SDWNLWVPLGILGIPTIWIALTYR
                      1   11   21   31

RC-M:
MADYQTIYTQIQARGPHITVSGEWGDNDRVGKPFYSYWLGKIGDAQIGPIYLGASGIAAFAFGSTAILIILFNMAAEVHFDPLQFFRQFFWLGLYPPKAQYGMGIPPLHDGGWWLMAGLFM
TLSLGSWWIRVYSRARALGLGTHIAWNFAAAIFFVLCIGCIHPTLVGSWSEGVPFGIWPHIDWLTAFSIRYGNFYYCPWHGFSIGFAYGCGLLFAAHGATILAVARFGGDREIEQITDRGT
AVERAALFWRWTIGFNATIESVHRWGWFFSLMVMVSASVGILLTGTFVDNWYLWCVKHGAAPDYPAYLPATPDPASLPGAPK

RC-H:
MYHGALAQHLDIAQLVWYAQWLVIWTVVLLYLRREDRREGYPLVEPLGLVKLAPEDGQVYELPYPKTFVLPHGGTVTVPRRRPETRELKLAQTDGFEGAPLQPTGNPLVDAVGPASYAERA
EVVDATVDGKAKIVPLRVATDFSIAEGDVDPRGLPVVAADGVEAGTVTDLWVDRSEHYFR YLELSVAGSARTALIPLGFCDVKKDKIVVTSILSEQFANVPRLQSRDQITLREEDKVSAYY
AGGLLYATPERAESLL

RC-L:
MALLSFERKYRVRGGTLIGGDLFDFWVGPYFVGFFGVSAIFFIFLGVSLIGYAASQGPTWDPFAISINPPDLKYGLGAAPLLEGGFWQAITVCALGAFISWMLREVEISRKLGIGWHVPLA
FCVPIFMFCVLQVFRPLLLGSWGHAFPYGILSHLDWVNNFGYQYLNWHYN PGHMSSVSFLFVNAMALGLHGGLILSVANPGDGDKVKTAEHENQYFRDVVGYSIGALSIHRLGLFLASNIF
LTGAFGTIASGPFWTRGWPEWWGWWLDIPFWS

RC-CytC:
MKQLIVNSVATVALASLVAGCFEPPPATTTQTGFRGLSMGEVLHPATVKAKKERDAQYPPALAAVKAEGPPVSQVYKNVKVLGNLTEAEFLRTMTAITEWVSPQEGCTYCHDENNLASEAK
YPYVVARRMLEMTRAINTNWTQHVAQTGVTCYTCHRGTPLPPYVRYLEPTLPLNNRETPTHVERVETRSGYVVRLAKYTAYSALNYDPFTMFLANDKRQVRVVPQTALPLVGVSRGKERRP
LSDAYATFALMMSISDSLGTNCTFCHNAQTFESWGKKSTPQRAIAWWGIRMVRDLNMNYLAPLNASLPASRLGRQGEAPQADCRTCHQGVTKPLFGASRLKDYPELGPIKA AAK

Extended Data Fig. 8 | Amino acid sequence of polypeptides in the LH1–RC complex from *Blc. viridis*. Black, genome sequence; red, protein sequence; blue, missing in protein sequence.

α-polypeptides

```
Blc. viridis (P04123)            MATEYRTASWKLWLILDPRRVLTALFVYLTVIALLI HFGLLSTDRLNWWEFQRGLPKAAS LVV--VPPAVG-
Rps. rubrum (Q2RQ24)             --------MWRIWQLFDPRQALVGLATFLFVLALLI HFILLSTERFNWLEGASTKPVQTSMVMPSS--DLAV
Rba. sphaeroides (Q3J1A4)        -----MSKFYKIWMIFDPRRVFVAQGVFLFLLAVMI HLILLSTPSYNWLEISAAKYNRVAVAE--------
Rba. capsulatus (P02948)         -----MSKFYKIWLVFDPRRVFVAQGVFLFLLAVLI HLILLSTPAFNWLTVATAKHGYVAAAQ--------
Phs. molischianum (Q9R4K5)       --------MWKIWTLYDPRRTLSGLFTFLTVLGLLI HFLLLSTDRFNWLDGAREAHNV-------------
Phs. molischianum LH2 (P97253)   --MSNPKDDYKIWLVINPSTWLPVIWIVATVVAIAV HAAVLAAPGFNWIALGAAKSAAK------------
Rps. palustris (Q6N9L4)          --------MWRIWLLFDPRRALVLLFVFLFGLAIII HFILLSTSRFNWLDGPRAAKAASI S-LPFTPPSMPV
Tch. tepidum (D2Z0P2)            -MFTMNANLYKIWLILDPRRVLVSIVAFQIVLGLLI HMIVLSTD-LNWLDDNIPVSYQALGKK---------
                                         :::* : :*   :        :.: :*  :*::   **
```

β-polypeptides

```
Blc. viridis  (P04124)            ----MADLKPSLTGLTEEEAKEFHGIFVTSTVLYLATAVIV HYLVWTARPWIAPIPKGWVNLEGVQSALS----YLV
Rps. rubrum  (P02950)             ---MADKNDLSFTGLTDEQAQELHAVYMSGLSAFIAVAVLA HLAVMIWRPWF------------------------
Rba. sphaeroides (Q2RQ23)         ---MAEVKQESLSGITEGEAKEFHKIFTSSILVFFGVAAFA HLLVWIWRPWV-PGPNGYSALETLTQTLT----YLS
Rba. capsulatus (P95673)          ------MAERSLSGLTEEEAIAVHDQFKTTFSAFIILAAVA HVLVWVWKPWF------------------------
Phs. molischianum (Q3J1A3)        ---MADKSDLGYTGLTDEQAQELHSVYMSGLWLFSAVAIVA HLAVYIWRPWF------------------------
Phs. molischianum LH2 (D2Z0P1)    -----MAEQKSLTGLTDDEAKEFHAIFMQSMYAWFGLVVIA HLLAWLYRPWL------------------------
Rps. palustris (Q6N9L5)           ------MSDGSISGLSEAEAKEFHSIFVTSFFLFIVVAVVA HILAWMWRPWL-PKATGYAMDSVHQLTSF--LC---
Tch. tepidum (O32409)             MANSFVRGGGTLSGLSESEAQEFHGIFVTSFISFIVVAIDA HFLAWKWRPWL-PGVKGYALLDNASTAAQSVLSTLV
                                           :*::: :*  .*   :      :   .  .*   .   :**.
```

Extended Data Fig. 9 | Amino acid sequence alignment of LH1 α- and β-polypeptides in LH1–RC core complexes from purple photosynthetic bacteria. All sequences have been aligned relative to the His residue that ligates BChls in the LH1 complexes. The α- and β-polypeptides of the *P. molischianum* LH2 complex are included for comparison. The sequence alignment was performed using CLUSTAL O v.1.2.4.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

Parameter	(EMDB-3951, PDB 6ET5)
Data collection and processing	
Magnification	130,000
Voltage (kV)	300
Electron exposure (e–/Å^2)	2.25 (45 e- on 20 frames)
Defocus range (μm)	-1.0 to -3.0
Pixel size (Å)	1.06
Symmetry imposed	c1
Initial particle images (no.)	267,762
Final particle images (no.)	166,816
Map resolution (Å) (global)	2.9
FSC threshold	0.143
Map resolution range (Å)	~2.5-3.5
Refinement	
Initial model used (PDB)	6ET5
Model resolution (Å)	2.9
FSC threshold	0.143
Model resolution range (Å)	~2.5-3.5
Map sharpening B factor (Å^2)	Estimated automatically using RELION 2.0*
Model composition	
Non-hydrogen atoms	31994
Protein residues	3492
Ligands	75
B factors (Å^2)	
Protein	RELION auto-estimated
Ligand	RELION auto-estimated
R.m.s. deviations (Refmac5)	
Bond lengths (Å)	0.01
Bond angles (°)	3.21
Validation	
wwwPDB	
Clashscore	27
Poor rotamers (%)	5.5
Ramachandran plot (COOT)	
Favored (%)	86.01
Allowed (%)	9.67
Disallowed (%)	4.32
***Refmac5*‡**	
FSC	0.89 (0.62)
R factor	0.26 (0.40)
Angle (rms)	3.21 (3.45)
Bond (rms)	0.01 (0.02)
Chiral (rms)	0.23 (0.31)
EMRinger score	3.34

*Data taken from ref. [50].

‡These results are calculated from a density map, in which electron density contributed by the surrounding belt of detergent was removed by masking. The results from the unmasked model are presented in parentheses.

ARTICLE

Structure of photosynthetic LH1–RC supercomplex at 1.9 Å resolution

Long-Jiang Yu[1], Michihiro Suga[1], Zheng-Yu Wang-Otomo[2] & Jian-Ren Shen[1]*

Light-harvesting complex 1 (LH1) and the reaction centre (RC) form a membrane-protein supercomplex that performs the primary reactions of photosynthesis in purple photosynthetic bacteria. The structure of the LH1–RC complex can provide information on the arrangement of protein subunits and cofactors; however, so far it has been resolved only at a relatively low resolution. Here we report the crystal structure of the calcium-ion-bound LH1–RC supercomplex of *Thermochromatium tepidum* at a resolution of 1.9 Å. This atomic-resolution structure revealed several new features about the organization of protein subunits and cofactors. We describe the loop regions of RC in their intact states, the interaction of these loop regions with the LH1 subunits, the exchange route for the bound quinone Q_B with free quinone molecules, the transport of free quinones between the inside and outside of the LH1 ring structure, and the detailed calcium-ion-binding environment. This structure provides a solid basis for the detailed examination of the light reactions that occur during bacterial photosynthesis.

Photosynthesis converts light energy from the Sun into biologically useful chemical energy, thereby sustaining virtually all life forms on Earth. The photosynthetic apparatus of purple photosynthetic bacteria is simple and robust, and has been studied extensively[1]. In most such bacteria, there are two types of light-harvesting complex, LH1 and LH2. Light energy is first absorbed by the peripheral LH2, then transferred via LH1 rapidly and efficiently to the reaction centre (RC) to drive the primary photochemical reactions. LH1 exists in all purple bacteria and surrounds the RC to form an integral membrane protein–pigment supercomplex (LH1–RC), consisting of 32–36 subunits with a total molecular weight of approximately 400 kDa.

The structures of Ca^{2+}-bound and Sr^{2+}/Ba^{2+}-substituted LH1–RC supercomplexes have been determined at resolutions of 3.0 Å and 3.3 Å, respectively, from the thermophilic purple sulfur bacterium *Thermochromatium tepidum*[2,3]. These structures showed that the RC is surrounded by 16 heterodimers of the LH1 αβ-subunits containing 32 bacteriochlorophyll (BChl) *a* and 16 spirilloxanthin molecules, forming a completely closed elliptical ring. This is different from the structures of both LH1–RC dimers from *Rhodobacter sphaeroides*[4] and LH1–RC monomers from *Rhodopseudomonas palustris*[5], determined by X-ray diffraction at resolutions of 8.0 Å and 4.8 Å, respectively. Both show incomplete ring structures, with the dimer having a PufX subunit and the monomer having the transmembrane helix protein W in the ring opening. Furthermore, 16 Ca^{2+}-binding sites were observed in the C-terminal-loop region of the thermophilic LH1 complex, which was considered to be the reason for the unusual redshift and enhanced thermal stability of the thermophilic LH1[6,7].

The resolution of the LH1–RC structure reported so far, however, has not been sufficient to reveal the detailed organization of many of the cofactors involved in the energy and electron transfer reactions within this supercomplex. Here we report the structure of LH1–RC from *Tch. tepidum* at 1.9 Å resolution, which reveals the detailed arrangement and organization of a large number of cofactors. On the basis of this high-resolution structure, we have examined the energy transfer from LH1 to RC, the quinone and proton channels, and the possible roles of Ca^{2+}. These results greatly advance our understanding of bacterial photosynthetic light reactions.

Overall structure

Compared with those from the previous structural determination[2], the quality of the LH1–RC crystals was improved considerably here by optimization of the detergent and other conditions of crystallization, enabling the structure to be resolved at 1.9 Å resolution (Extended Data Figs. 1 and 2, Extended Data Table 1). The space group of the crystal obtained, *C2*, was the same as that reported in the previous study[2] (Extended Data Fig. 1c, d). However, the unit-cell dimensions were much shorter than before (Extended Data Table 1), leading to a more compact packing and a much lower solvent content of 55% (compared to 65% for the previous crystals[2]). This could be a major reason for the substantial improvement in the crystal resolution.

The overall structure of the LH1–RC complex is similar to that previously determined at 3.0 Å resolution[2] (Fig. 1). However, the root mean square deviation (r.m.s.d.) between Cα atoms of the two structures is 1.68 Å; this relatively large value mainly arises from deviations in some regions of the RC subunits and in the N and C termini of the LH1 subunits. We identified a number of lipid and detergent molecules in the 'gap region' between RC and LH1 (Fig. 1c), resulting in a rather crowded gap region in comparison with that from the previous structure, which was relatively empty (Extended Data Fig. 3). The RC contains four protein subunits, four BChls *a*, two bacteriopheophytins, one Mg^{2+} ion, one Fe^{3+} ion, one spirilloxanthin molecule, one molecule of menaquinone-8 (MQ8) and five molecules of ubiquinone-8 (UQ8). The LH1 contained 16 pairs of LH1 αβ-subunits, 32 BChls *a*, 16 spirilloxanthin molecules and 16 Ca^{2+} ions (Fig. 1, Extended Data Table 2). In addition, nearly 1,000 water molecules were found in the supercomplex, mostly distributed in the hydrophilic surfaces of the cytoplasm and periplasm (Fig. 1b). LH1–RC is elliptical in shape, as has been reported previously[2], but the distances between the BChls of the RC and the closest LH1 BChls are almost equal across the whole structure (Fig. 1d). This could ensure an efficient energy transfer from LH1 to RC.

The r.m.s.d. values between Cα atoms of the new RC structure and the RC-only structure at 2.2 Å resolution[8], or the RC structure at 3.0 Å resolution[2], were 3.19 Å and 1.98 Å, respectively, due to apparent differences in some regions of the RC structures (Extended Data Figs. 3b, 4).

---

[1]Research Institute for Interdisciplinary Science, Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan. [2]Faculty of Science, Ibaraki University, Mito, Japan.
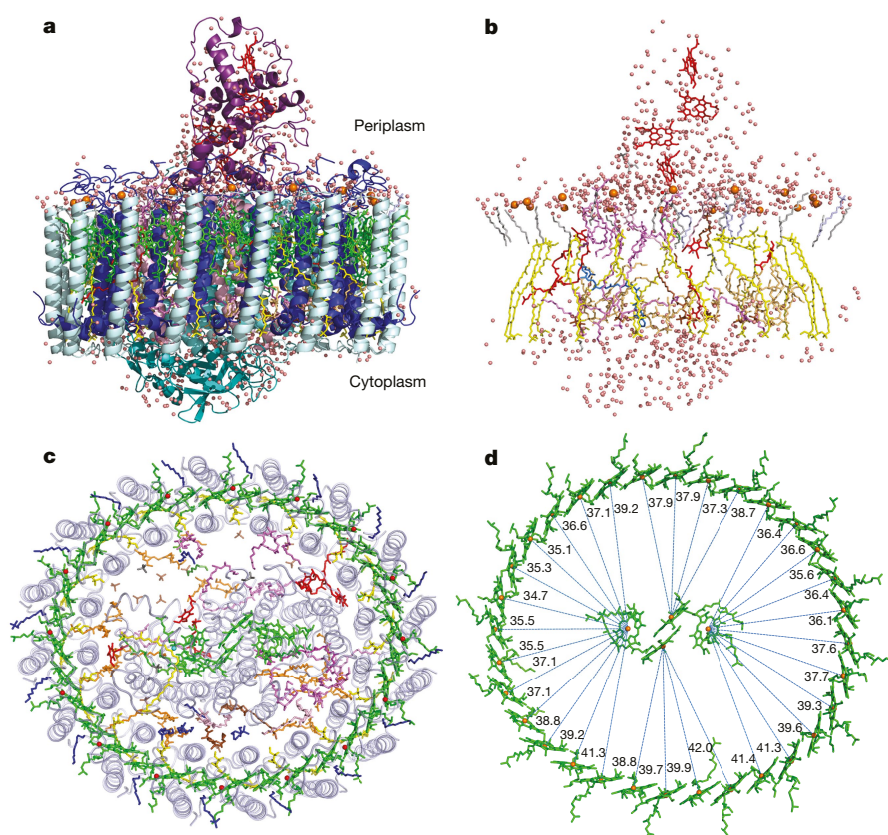*e-mail: shen@cc.okayama-u.ac.jp

Fig. 1 | Architecture of LH1–RC from *Tch. tepidum* at a resolution of 1.9 Å. a, View from the direction parallel to the membrane plane. LH1 α-subunit, blue; LH1 β-subunit, light cyan; C subunit, purple. **b,** Arrangement of the cofactors and water molecules, with the same view as in **a**. **c,** Arrangement of the cofactors, viewing perpendicular to the membrane. Protein subunits are depicted in light grey. **d,** Distances of the closest BChl pairs between RC and the surrounding LH1 ring. Colour codes for cofactors: BChls, green; spirilloxanthin, yellow; Ca^{2+} ions, orange spheres; water molecules, pale pink dots.

The C-terminal loop regions of all LH1 α-apoproteins, which have poor electron densities in the previous 3.0 Å structure[2], show relatively large deviations (Extended Data Fig. 4c, d). Because these terminal regions are exposed to the surface of the membrane and partly flexible, as shown by their high B-factors (Extended Data Table 3), we adjusted the lengths of the α- and β-apoproteins with confidence based on the electron density map. Furthermore, the positions and coordination patterns of the calcium ions were identified unambiguously. Next, we describe features found in the present high-resolution structure that are unique and important for the function of this supercomplex.

Unique features in the structure of RC

Three major differences were found between the present intact RC and the previous isolated RC structures. These include the cytochrome (Cyt) subunit N-terminal region and its loop region (residues 172–196), and a loop region in the H subunit (residues 44–58) (Fig. 2a, c, d). The Cyt subunit is located at the periplasmic side, and has been reported to be a lipoprotein in *Blastochloris viridis* with its N-terminal cysteine linked to a diglyceride via a thioether bond[9–11]. In the previous structure[2], this region was assigned incorrectly because of the lower resolution, or possibly because of damage by X-ray radiation that could break the thioether bond[11]. In the present structure, the electron-density map showed three partial acyl chains attached to Cys23 of the Cyt subunit. Among these, one is a single chain and another one is clearly branched (Fig. 2b). This suggests that the N-terminal cysteine is triacylated with *N*-acyl and *S*-diacylglycerol in a manner similar to that of an outer membrane protein[12], and these fatty acids anchor the subunit in the membrane. Beyond the seventh carbon from the carbonyl carbon, all aliphatic tails are disordered—presumably due to their flexibility—and could interact

with UQ8, which was found in a location appropriate for its forthcoming exchange with Q_B. In addition, the loop region (172–196) of Cyt showed a large deviation from that of the isolated RCs[8,9], and appears to interact with the neighbouring LH1 only in this conformation (Fig. 2c). A Mg^{2+} ion was found in its vicinity (Fig. 2c, e), which may reduce the flexibility of this long loop.

The loop region (44–58) of the H subunit was traced unambiguously in the present structure (Fig. 2d), and interacts with the neighbouring LH1 β-polypeptide at the cytoplasmic side. This is largely different from the isolated RCs and the previous LH1–RC complex. In the recently reported structure of the RC from *Blc. viridis*[9] at 1.92 Å resolution, these residues are located in the crystal lattice contact region, which is different from their location in previous structures[11,13]. The interactions with the LH1 polypeptides therefore seem to be required to maintain this region in the correct configuration.

Quinones and lipids

One molecule of MQ8 and five molecules of UQ8 were found in the LH1–RC supercomplex (Fig. 3a), consistent with the results of biochemical analysis of the same sample[14]. Among these quinones, MQ8 and one of the UQ8 molecules function as the primary (Q_A) and secondary (Q_B) quinone acceptors, respectively, with similar binding sites to those reported previously[2]. The additional UQ8 molecules were found to be distributed over the RC and in the gap region between the RC and LH1. In particular, one of the UQ8 molecules was located in a position close to the isoprenoid tail of Q_B near to the periplasmic surface (Fig. 3a), with its head oriented in the same direction as that of Q_B, suggesting that this quinone is in a position appropriate for the exchange of Q_B after its double reduction and protonation. This is supported by the fact that, whereas the head of Q_B is hydrogen bonded to
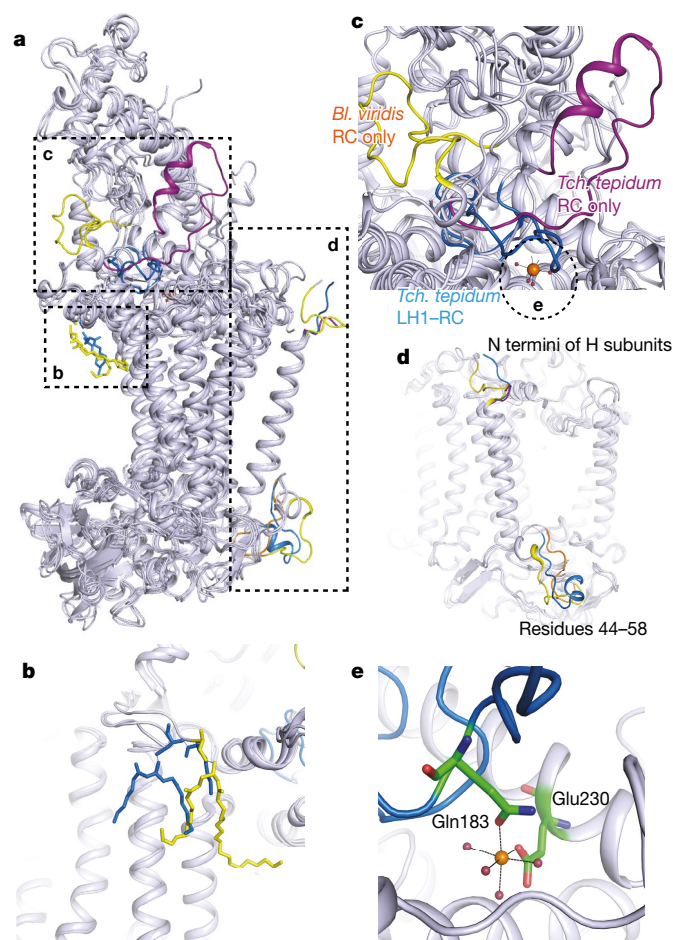
Fig. 2 | Differences between the isolated and intact RC core structures. **a**, Superposition of the RC structures from *Tch. tepidum* (1EYS; purple), *Rba. sphaeroides* (2J8C; orange), *Blc. viridis* (3T6E; yellow) and the present structure (blue). Regions similar in structure are coloured in grey, whereas the three regions with large differences are boxed and coloured differently. The boxed areas are enlarged in **b**, **c** and **d**, as indicated. **b**, Triacylation of the Cyt N-terminal cysteine. **c**, The loop region of Cyt (residues 172–196) including the Mg^{2+}-binding site, which is circled here and enlarged in **e**. **d**, The N-terminal region and residues 44–58 of the H subunit. **e**, The Mg^{2+}-binding site in the Cyt subunit.

the L-subunit residue His199, L-Ser232, L-Ile233 and L-Gly234, the head of the second UQ8 was not hydrogen bonded to any residues, and its isoprenoid tail was not visible, presumably owing to its high flexibility (Fig. 3a). The cavity containing the head of this UQ8 was formed by L-Met183, L-Leu184, L-Ser187, L-Trp272, M-Ile179, as well as accessory BChl bound to the M subunit.

Three UQ8s are located in the gap region between LH1 and RC. Among these, two are close to the LH1 ring, and the isoprenoid tail of one of these two UQ8s was found to be inserted into a space between the LH1 α- and β-subunits (Fig. 3b, c), which suggests that it is undergoing transport between the inside and outside of the ring through the possible exchange channel between the LH1 subunits. This channel is close to the cytoplasmic side of the membrane at the same level as the head of Q_B, and has been suggested in previous studies[2,15], but this present result provides direct evidence for the transport of quinones through such channels. This channel is hydrophobic and is surrounded by Val20, Ser23, Ile24 and Phe27 from an α-subunit of LH1 on one side, and Leu21, Val22, Val25 and Ile29 from another α-subunit on the opposite side (Fig. 3c). In addition, the spirilloxanthin and the phytol tail of BChl bound to the LH1 α-subunit are near to the exit of the channel, and may contribute to the formation of the hydrophobic channel exit.
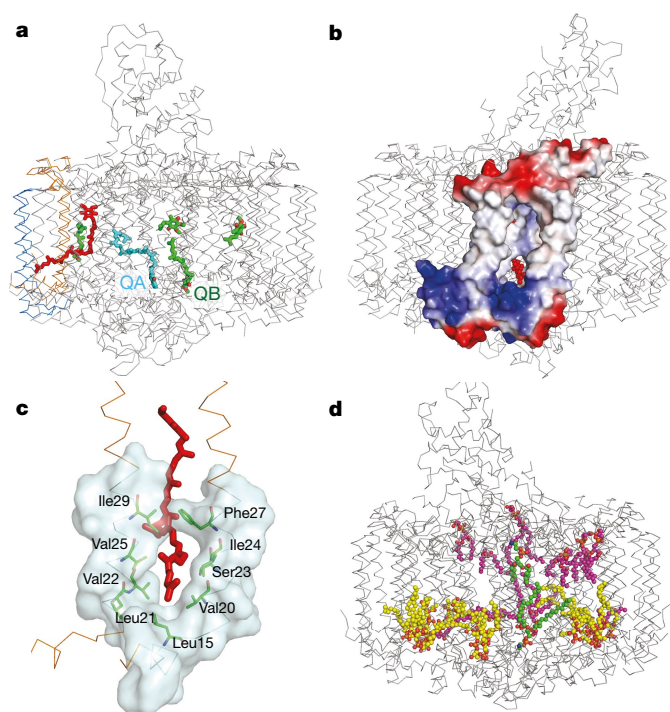

Fig. 3 | Distribution of quinones and lipids in LH1–RC. a, Overall distribution of six quinones in the present structure. UQs, green; MQ, cyan; UQ (with its tail inserted in the channel between the LH1 α- and β-subunits), red. The two α- and two β-subunits of LH1 that form the channel for transport of the UQ are shown in orange and blue, respectively. **b**, **c**, The quinone-exchange channel between the LH1 α- and β-subunits. **d**, Distribution of lipids from a side view of LH1–RC. The lipid molecules are shown in space-filling mode. Oxygen, red; nitrogen, blue; PEF carbons, green; CDL carbons, yellow; phosphatidylglycerol carbons, magenta; proteins, grey.

Extensive hydrogen-bonding networks were found in the H subunit that connects Q_B to the cytoplasmic surface (Extended Data Fig. 5), owing to the large number of water molecules. One of the major hydrogen-bonding networks is approximately perpendicular to the membrane surface, and may serve as a proton-transfer channel to connect Q_B to the aqueous phase[16]. We also found a water cluster parallel to the membrane plane, similar to that found in *Rba. sphaeroides* RC[16–18]. These hydrogen-bond networks support the idea that there are multi-entry proton uptake networks for the protonation of Q_B[19].

Among the 21 lipids identified, nine are tentatively assigned to cardiolipins (CDLs), ten to phosphatidylglycerols and two to phosphatidylethanolamines (PEFs) (Fig. 3d, Extended Data Table 2). This number is much larger than for the RC-only structure in which only 1–2 lipids were found, presumably due to the loss or disorder of the lipid molecules upon solubilization of the RC-only complex, or replacement by the detergents used. The distribution of the lipids is asymmetric: all of the CDLs were found at the cytoplasmic side with their head groups localized at the surface of the membrane, whereas PEFs and phosphatidylglycerols were located at both the cytoplasmic and the periplasmic sides. All of the three lipids that have previously been reported at 3.0 Å resolution[2] were confirmed, but with some modifications: the PEF and one of the two phosphatidylglycerols were reassigned as CDL and PEF, respectively.

Interactions of LH1–RC and interactions among LH1

LH1 and RC interact at both the periplasmic and the cytoplasmic sides, either directly (Fig. 4a, b) or indirectly (for example, through lipids; Extended Data Table 4). As discussed, the newly built loop region
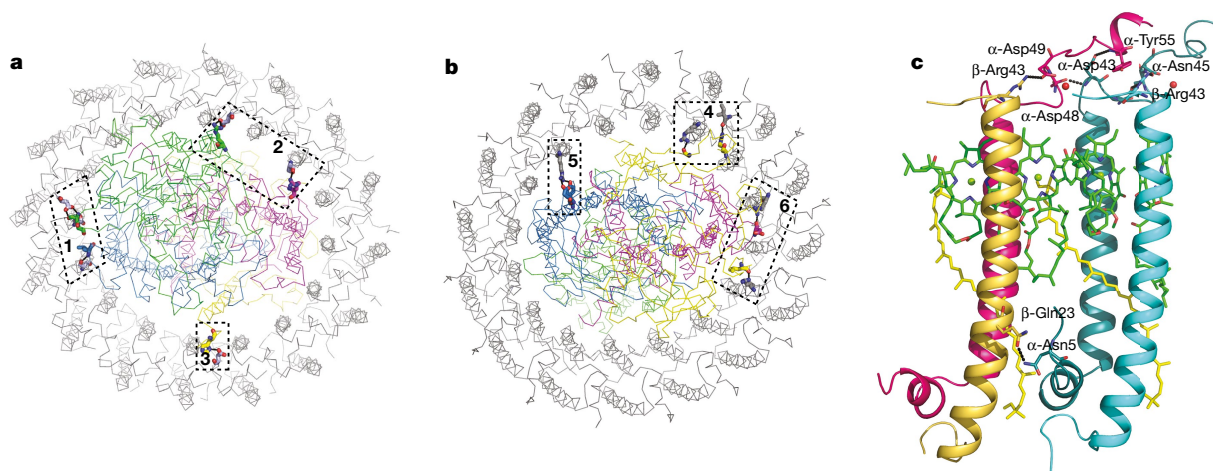
Fig. 4 | Interactions between LH1 and RC, and among LH1. For clarity, only protein–protein interactions are depicted. **a**, **b**, Interaction sites (boxed) between LH1-α and RC at the periplasmic side (**a**) and the cytoplasmic side (**b**). The numbers 1–6 in the boxed areas correspond to panels a–f in Extended Data Fig. 6. L subunit, magenta; M subunit, blue; C subunit, green; H subunit, yellow; LH1 α- and β-subunits, light grey. **c**, Interactions between adjacent LH1 α–α and α–β subunits. LH1 α-subunits, magenta and dark cyan; LH1 β-subunits, dark yellow and cyan; BChl, green; spirilloxanthin, light yellow.

(172–196) of Cyt forms two hydrogen bonds between C-Ser176, C-Gly177 and the neighbouring α-Asp48, α-Ser41; this may stabilize this region of the Cyt subunit, indicating that this conformation represents its intact state. In addition, M-Leu109, C-Arg47, L-Arg85, and H-His7 interact with the neighbouring LH1 α-Ser41 or α-Asp48 at the periplasmic side (Extended Data Fig. 6a–c). Importantly, α-Ser41 and α-Asp48 are two of the main residues involved in the interaction between LH1 and RC subunits or lipids at the periplasmic side (Extended Data Table 4). At the cytoplasmic side, two arginine residues (Arg18 and Arg19) located at the beginning of the α-subunit helices have a major role in interacting with the RC subunits or lipids, and other residues (Il414, Asp16 and Ser23) are also involved in interactions with the RC and lipids (Fig. 4b, Extended Data Fig. 6d–f, Extended Data Table 4). These two arginines, together with some other arginines and lysines from the RC and α-Lys10 or β-Lys15, form a positively charged layer at the membrane surface. This might interact with the phosphate group of the lipids, thereby strengthing the association of LH1 with the RC.

Extensive interactions between the LH1 αβ-heterodimers are found near to the Ca²⁺-binding site in the C-terminal domain on the periplasmic side, especially in the region of residues α-42–49 (Fig. 4c). This region forms a characteristic short turn structure at the surface of the membrane; this occurs only in *Tch. tepidum*, because a residue at the position 43 is deleted in this organism in comparison with other photosynthetic bacteria[20]. This characteristic structure enables α-Asp43 to form hydrogen bonds with α-Asp48, α-Ser54, α-Tyr55 and α-Gln56 from the neighbouring α-polypeptide. In addition, α-Asn45 is also a key residue because it is conserved in almost all purple bacteria, and forms extensive interactions with its neighbouring subunits. β-Arg43 is also highly conserved and is involved in additional interactions with its neighbouring subunits (Extended Data Table 4). Another strong hydrogen bond was found between β-Pro44 and α-Tyr55. Conversely, only one strong hydrogen bond was found between β-Leu46 and (n − 1) β-Tyr42 at the membrane surface for the β–β interactions.

In contrast to the extensive interactions in the C-terminal region on the periplasmic side, there are fewer interactions in the N-terminal region on the cytoplasmic side. β-Thr7 is hydrogen bonded to α-Leu13 and α-Trp12, and β-Asp11 interacts with α-Tyr9 and α-Lys10 (Fig. 4c). Taken together, the interactions in both the N-terminal and C-terminal regions ensure a tight connection of the LH1 αβ-apoproteins as well as a joint coordination to BChls *a*, carotenoids and Ca²⁺ ions (Fig. 5a, b), which results in a robust, closed concentric elliptical ring structure.

Ca²⁺-binding sites

One of the notable features of the thermophilic LH1–RC is its binding of 16 Ca²⁺ ions in the LH1 subunits. We identified all of the ligands for Ca²⁺ unambiguously: the side chain of α-Asp49, the carbonyl oxygens of α-Trp46, α-Ile51, (n + 1) β-Trp45, and two water molecules, giving rise to a six-coordinate structure (Fig. 5c, d). Three out of the four coordinating residues are hydrophobic. This might be due to the fact that this binding site is located just at the periplasmic surface of the membrane, which may contribute to the weak binding of Ca²⁺ and render it easily exchangeable with other divalent cations[21,22].

The Ca²⁺-binding site is located in the C-terminal region of both αβ subunits, which contributes to the tight connection of the two LH1 subunits (Fig. 5). This is in accordance with the result of Fourier-transform infrared spectroscopy measurements, which showed that the
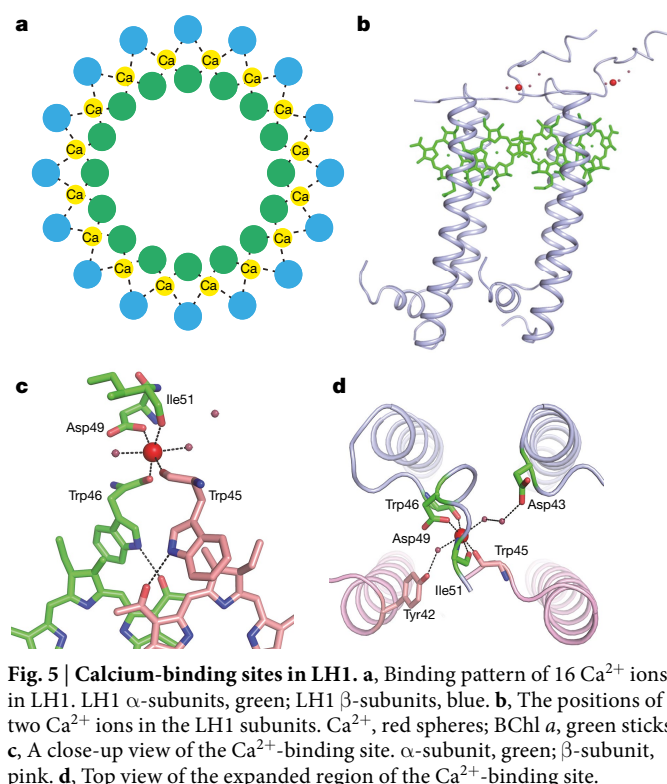

Fig. 5 | Calcium-binding sites in LH1. a, Binding pattern of 16 Ca²⁺ ions in LH1. LH1 α-subunits, green; LH1 β-subunits, blue. **b**, The positions of two Ca²⁺ ions in the LH1 subunits. Ca²⁺, red spheres; BChl *a*, green sticks. **c**, A close-up view of the Ca²⁺-binding site. α-subunit, green; β-subunit, pink. **d**, Top view of the expanded region of the Ca²⁺-binding site.

binding of Ca^{2+} reduces the conformational flexibility of LH1–RC[23]. The structural stability induced by the binding of Ca^{2+} may therefore contribute to the thermophilic stability of LH1 as well as the redshift of the absorption peak, two unique features of *Tch. tepidum*. These results are in agreement with those of recent spectral measurements[24–26].

The unique binding of Ca^{2+} may be related to the deletion of the residue α-43 in *Tch. tepidum*[20], as the insertion of an alanine into this site has been shown to disrupt the Ca^{2+} binding of the thermophilic LH1, leading to a blueshift in its absorption[27]. This provides support for the Ca^{2+}-binding environment revealed in the present study and for its functional importance.

Differences were also found in the BChls of LH1 between the present and the previous structures. The imidazole ring of β-His36, a direct ligand for β-BChl, is rotated by about 45° in most cases, and the porphyrin plane of the β-BChl is rotated by about 10° along its Q_y axis (the axis connecting pyrrole rings I–III). These changes resulted in a more parallel orientation of the neighbouring BChls, giving rise to a stronger coupling between the adjacent BChls.

The novel features of LH1–RC revealed by this high-resolution structural analysis—including the unique conformations of several loop regions of the Cyt and H subunits and their interaction sites with LH1, the location of additional ubiquinones, the presence of water clusters that form hydrogen-bonding networks, and the detailed Ca^{2+}-binding site—provide important information regarding the energy transfer between LH1 and RC, the shuttling of ubiquinone through LH1 to the cytochrome bc_1 complex and proton transfer to Q_B, and the roles of Ca^{2+} in the redshift and high thermostability of LH1–RC from *Tch. tepidum*.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0002-9.

1. Cogdell, R. J. & Roszak, A. W. Structural biology: the purple heart of photosynthesis. *Nature* **508**, 196–197 (2014).
2. Niwa, S. et al. Structure of the LH1–RC complex from *Thermochromatium tepidum* at 3.0 Å. *Nature* **508**, 228–232 (2014).
3. Yu, L. J., Kawakami, T., Kimura, Y. & Wang-Otomo, Z. Y. Structural basis for the unusual Q_y red-shift and enhanced thermostability of the LH1 complex from *Thermochromatium tepidum*. *Biochemistry* **55**, 6495–6504 (2016).
4. Qian, P. et al. Three-dimensional structure of the *Rhodobacter sphaeroides* RC-LH1-PufX complex: dimerization and quinone channels promoted by PufX. *Biochemistry* **52**, 7575–7585 (2013).
5. Roszak, A. W. et al. Crystal structure of the RC-LH1 core complex from *Rhodopseudomonas palustris*. *Science* **302**, 1969–1972 (2003).
6. Kimura, Y. et al. Calcium ions are involved in the unusual red shift of the light-harvesting 1 Q_y transition of the core complex in thermophilic purple sulfur bacterium *Thermochromatium tepidum*. *J. Biol. Chem.* **283**, 13867–13873 (2008).
7. Kimura, Y., Yu, L. J., Hirano, Y., Suzuki, H. & Wang, Z. Y. Calcium ions are required for the enhanced thermal stability of the light-harvesting-reaction center core complex from thermophilic purple sulfur bacterium *Thermochromatium tepidum*. *J. Biol. Chem.* **284**, 93–99 (2009).
8. Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T. & Miki, K. Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from *Thermochromatium tepidum*: thermostability and electron transfer. *Proc. Natl Acad. Sci. USA* **97**, 13561–13566 (2000).
9. Roszak, A. W. et al. New insights into the structure of the reaction centre from *Blastochloris viridis*: evolution in the laboratory. *Biochem. J.* **442**, 27–37 (2012).
10. Weyer, K. A., Schafer, W., Lottspeich, F. & Michel, H. Cytochrome subunit of the photosynthetic reaction center from *Rhodopseudomonas viridis* is a lipoprotein. *Biochemistry* **26**, 2909–2914 (1987).
11. Wöhri, A. B. et al. Lipidic sponge phase crystal structure of a photosynthetic reaction center reveals lipids on the protein surface. *Biochemistry* **48**, 9831–9838 (2009).
12. Kulathila, R., Kulathila, R., Indic, M. & van den Berg, B. Crystal structure of *Escherichia coli* CusC, the outer membrane component of a heavy metal efflux pump. *PLoS ONE* **6**, e15610 (2011).
13. Li, L. et al. Nanoliter microfluidic hybrid method for simultaneous screening and optimization validated with crystallization of membrane proteins. *Proc. Natl Acad. Sci. USA* **103**, 19243–19248 (2006).
14. Kimura, Y. et al. Characterization of the quinones in purple sulfur bacterium *Thermochromatium tepidum*. *FEBS Lett.* **589**, 1761–1765 (2015).
15. Wang-Otomo, Z.-Y. in *Solar to Chemical Energy Conversion* (eds Sugiyama, M. et al.) 379–390 (Springer International Publishing, Cham, 2016).
16. Stowell, M. H. B. et al. Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron-proton transfer. *Science* **276**, 812–816 (1997).
17. Fritzsch, G., Kampmann, L., Kapaun, G. & Michel, H. Water clusters in the reaction centre of *Rhodobacter sphaeroides*. *Photosynth. Res.* **55**, 127–132 (1998).
18. Abresch, E. C. et al. Identification of proton transfer pathways in the X-ray crystal structure of the bacterial reaction center from *Rhodobacter sphaeroides*. *Photosynth. Res.* **55**, 119–125 (1998).
19. Krammer, E. M., Till, M. S., Sebban, P. & Ullmann, G. M. Proton-transfer pathways in photosynthetic reaction centers analyzed by profile hidden Markov models and network calculations. *J. Mol. Biol.* **388**, 631–643 (2009).
20. Rücker, O., Köhler, A., Behammer, B., Sichau, K. & Overmann, J. *Puf* operon sequences and inferred structures of light-harvesting complexes of three closely related *Chromatiaceae* exhibiting different absorption characteristics. *Arch. Microbiol.* **194**, 123–134 (2012).
21. Kimura, Y., Inada, Y., Yu, L. J., Wang, Z. Y. & Ohno, T. A spectroscopic variant of the light-harvesting 1 core complex from the thermophilic purple sulfur bacterium *Thermochromatium tepidum*. *Biochemistry* **50**, 3638–3648 (2011).
22. Kimura, Y. et al. Metal cations modulate the bacteriochlorophyll–protein interaction in the light-harvesting 1 core complex from *Thermochromatium tepidum*. *Biochim. Biophys. Acta* **1817**, 1022–1029 (2012).
23. Jakob-Grun, S., Radeck, J. & Braun, P. Ca^{2+}-binding reduces conformational flexibility of RC-LH1 core complex from thermophile *Thermochromatium tepidum*. *Photosynth. Res.* **111**, 139–147 (2012).
24. Ma, F., Yu, L. J., Wang-Otomo, Z. Y. & van Grondelle, R. The origin of the unusual Q_y red shift in LH1–RC complexes from purple bacteria *Thermochromatium tepidum* as revealed by Stark absorption spectroscopy. *Biochim. Biophys. Acta* **1847**, 1479–1486 (2015).
25. Ma, F., Yu, L. J., Wang-Otomo, Z. Y. & van Grondelle, R. Temperature dependent LH1 → RC energy transfer in purple bacteria *Tch. tepidum* with shiftable LH1-Q_y band: a natural system to investigate thermally activated energy transfer in photosynthesis. *Biochim. Biophys. Acta* **1857**, 408–414 (2016).
26. Ma, F. et al. Metal cations induced αβ-BChl *a* heterogeneity in LH1 as revealed by temperature-dependent fluorescence splitting. *ChemPhysChem* **18**, 2295–2301 (2017).
27. Nagashima, K. V. P. et al. Probing structure–function relationships in early events in photosynthesis using a chimeric photocomplex. *Proc. Natl Acad. Sci. USA* **114**, 10906–10911 (2017).

Author contributions J.-R.S. and L.-J.Y. conceived the project; L.-J.Y. prepared the samples with the help of Z.-Y.W.-O.; L.-J.Y. grew the crystals under the supervision of J.-R.S.; L.-J.Y. and M.S. collected the diffraction data and analysed the structure; L.-J.Y. and J.-R.S. wrote the manuscript; and all authors contributed to the discussion and improvement of the manuscript.

METHODS

No statistical method was used to predetermine the sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Purification and crystallization. *Tch. tepidum* cells were grown in a growth chamber (BiOTRON, LH-410PFP-SP, NK System,) at 49 °C for 7 days. The light illumination was provided by LED lamps specified for plant growth, which have emission peaks at around 450 nm and 645 nm, respectively, at a light intensity of $30 \, \mu E \, m^{-2} \, s^{-1}$. The bacterial cells grown under these conditions appeared to have a larger ratio of LH1–RC/LH2 according to the absorption spectra, which suggests that there are more LH1–RC in the same amount of wet bacterial cells. LH1–RC complex was purified as described previously[2,28] with slight modifications. The final LH1–RC samples with a ratio of A_{915}/A_{280} greater than 2.20 were collected and precipitated by addition of polyethylene glycol 1,450 to a final concentration of 13% (w/v), and then suspended in 20 mM MES (pH 6.2) containing 3.4% *n*-octyl-phosphocholine (OPC) to a concentration of 20 mg protein per ml. Crystallization was performed by a microbatch-under-oil method, in which 2 μl of the above protein solution was mixed with an equal volume of the precipitant solution containing 50 mM MES (pH 6.2), 50 mM CaCl₂, 10 mM MgCl₂, 3.4% OPC and 26% polyethylene glycol 1,450. The crystals grew to sizes of $0.3 \times 0.4 \times 0.05 \, mm^3$ to $0.4 \times 0.8 \times 0.2 \, mm^3$ in 10 days at 20 °C (Extended Data Fig. 1a), and were then transferred into a 10 μl cryoprotectant solution containing 50 mM MES (pH 6.2), 3.4% OPC, 30% polyethylene glycol 1,450, 50 mM CaCl₂ and 15% glycerol, and flash-frozen immediately in a nitrogen stream.

Data collection. X-ray diffraction experiments were carried out at beamlines BL41XU of SPring-8 and BL1A of the Photon Factory (Japan). The highest-resolution diffraction data used for structural analysis was collected at BL41XU of SPring-8. The wavelength of X-rays used was 1.0 Å and the beam size was $35 \times 22 \, \mu m^2$. The diffraction images were recorded with a Pilatus 6 M detector, and the crystals were rotated by 0.1° in a helical manner. A total of 5,400 images covering a rotation angle of 540° were collected. The photon flux of the beamline used was 6.8×10^{11} photons per second (after attenuation by 0.75-mm thick aluminium), and the exposure time was 0.1 s for each diffraction image. The diffraction data was processed, integrated and scaled using the XDS Program Package (version October 15, 2015)[29], and the reflection data statistics are summarized in Extended Data Table 1.

Structure refinement. The initial structure of the LH1–RC complex was solved by the molecular replacement method using the Phaser program in PHENIX (version 1.12-2829)[30]. The structure of LH1–RC previously determined at 3.0 Å resolution from *Tch. tepidum* (PDB code: 3WMM) was used as the search model, with the Ca²⁺ ions, lipid and solvent molecules omitted. Five per cent of reflections were used for the free *R* factor calculation in the structure refinement. The initial model was subjected to rigid-body and restrained refinements successively in a resolution range of 50–2.0 Å. Incorporation of cofactors, lipid and detergent molecules and model modification were performed using COOT (version 0.8.2)[31]. For the assignment of lipid molecules, positions of the phosphorous atoms in lipids were confirmed either by the strong electron density interacting with the positively charged amino acid residues or the peaks found in the anomalous map. The lipids were assigned, on the basis of the electron density of the polar head group, as CDLs when they were connected to each other, and as PEFs and phosphatidylglycerols when they interact with the negatively charged amino acids and neutral groups, respectively. Positional and isotropic parameters were refined in the resolution range of 50–1.9 Å. After solvent molecules were included in the model, TLS (translation, libration, screw) refinement was performed, and the final model was refined to $R_{work} = 18.15\%$ and $R_{free} = 21.52\%$, with 98.42% of residues in the favoured Ramachandran region, 1.51% in the allowed region and 0.08% in the outliers. The relatively high *R* values may be attributed to blurred electron densities at the terminal regions of the LH1 polypeptides, especially the N terminus, resulting in higher B-factors in these regions. In addition, some residual densities in the gap region between RC and the LH1 ring were not modelled, and they may be flexible fragments of lipids and quinones. The refinement statistics are listed in Extended Data Table 1, and the quality of the structure was analysed by using PROCHECK[32]. Figures were generated with the PyMOL program[33].

Reporting Summary. Further information on experimental design is available in the Reporting Summary linked to this paper.

Data availability. Atomic coordinates and structure factors for the reported crystal structure have been deposited in the Protein Data Bank under accession number 5Y5S.

28. Suzuki, H. et al. Purification, characterization and crystallization of the core complex from thermophilic purple sulfur bacterium *Thermochromatium tepidum*. *Biochim. Biophys. Acta* **1767**, 1057–1063 (2007).
29. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
30. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
31. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
32. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
33. The PyMOL Molecular Graphics System v.1.8 (Schrödinger, 2015).
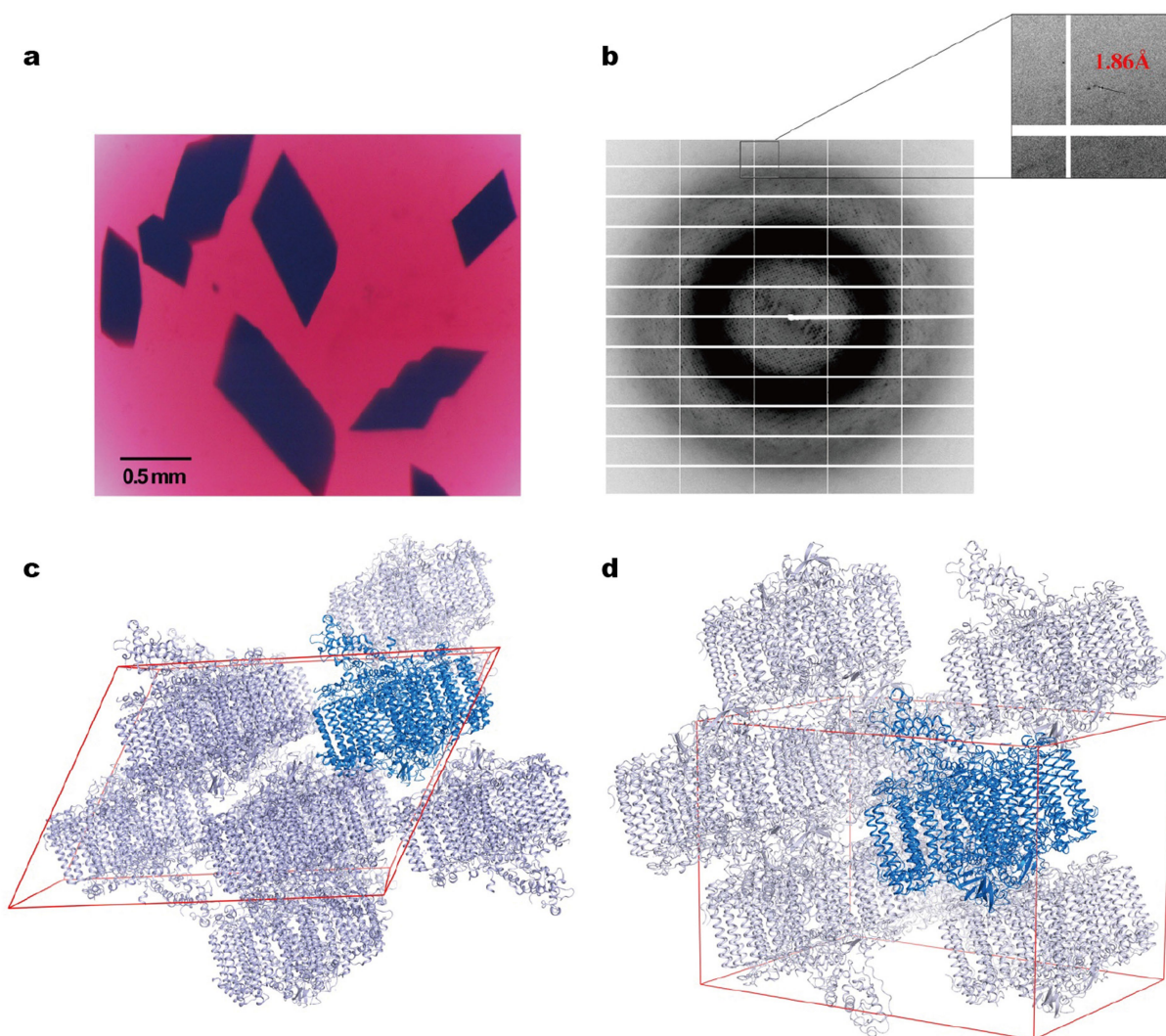
Extended Data Fig. 1 | Quality of the LH1–RC crystal and its packing pattern. a, An image of the LH1–RC crystals obtained in the present study. These crystals were obtained reproducibly under the present crystallization conditions. **b**, A typical diffraction image of the LH1–RC crystal taken at BL41XU of SPring-8, Japan, with a wavelength of 1.0 Å at 100 K. This diffraction image was obtained reproducibly with many crystals tested. **c**, **d**, Packing patterns of the previous (**c**) and the present crystal (**d**).
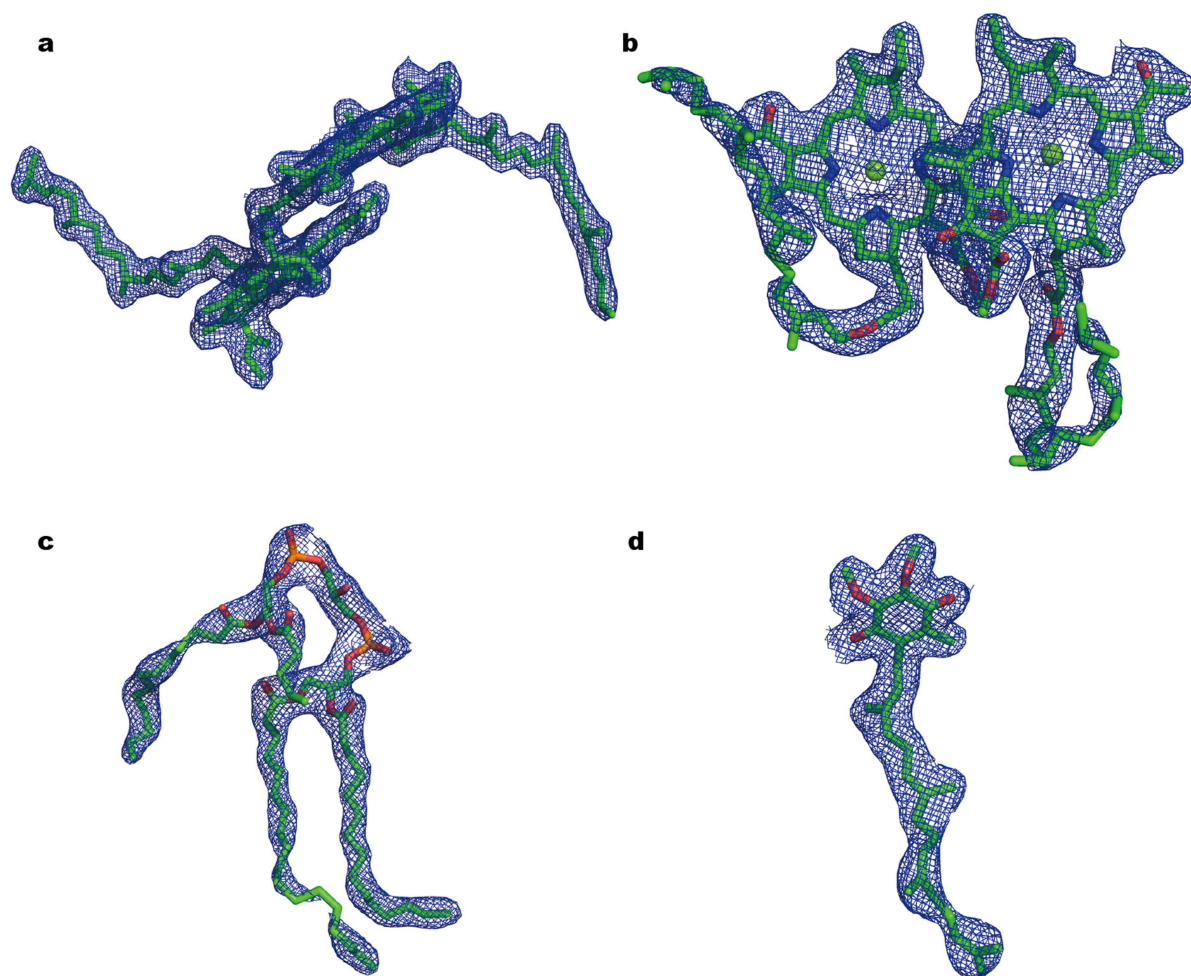
Extended Data Fig. 2 | Close-up views of the electron density maps for some of the cofactors of LH1–RC. The blue mesh represents the $2F_o - F_c$ map contoured at 1.0σ, taken at a wavelength of 1.0 Å and analysed to 1.9 Å resolution. **a–d**, The special-pair BChls (**a**), one pair of the LH1 BChls (**b**), one of the CDL (**c**) and the Q_B molecule (**d**).
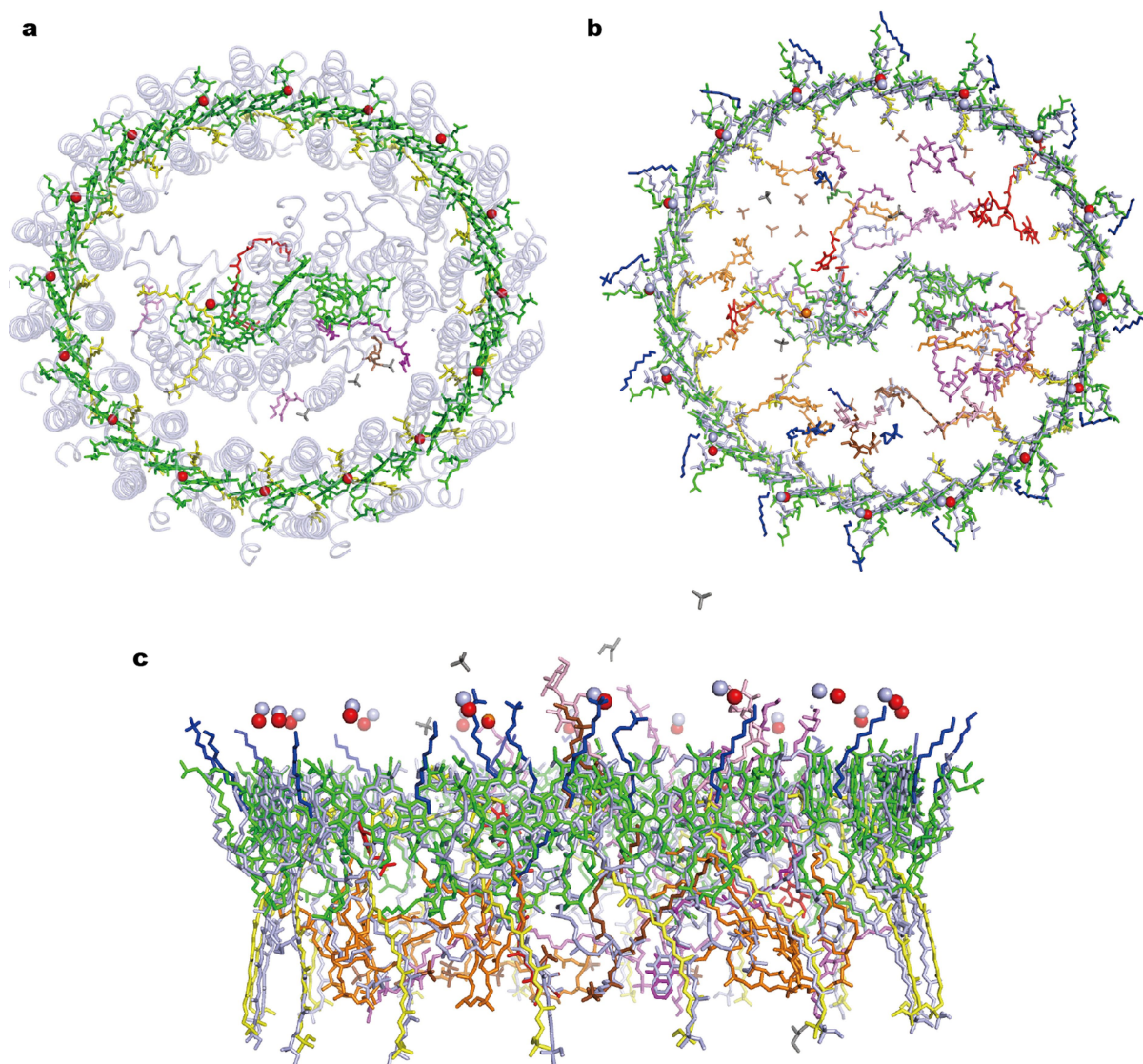
Extended Data Fig. 3 | Comparison of the arrangement of the cofactors between the previous and present structures. a, Arrangement of the cofactors in the previous 3.0 Å structure, with a view from the top of the membrane. **b**, Superposition of the cofactors between the previous 3.0 Å and present 1.9 Å structures. **c**, The same as **b**, viewing from the side of the membrane. In **b** and **c**, the cofactors revealed in the present 1.9 Å structures are coloured differently, whereas those in the previous 3.0 Å structures are depicted in grey.
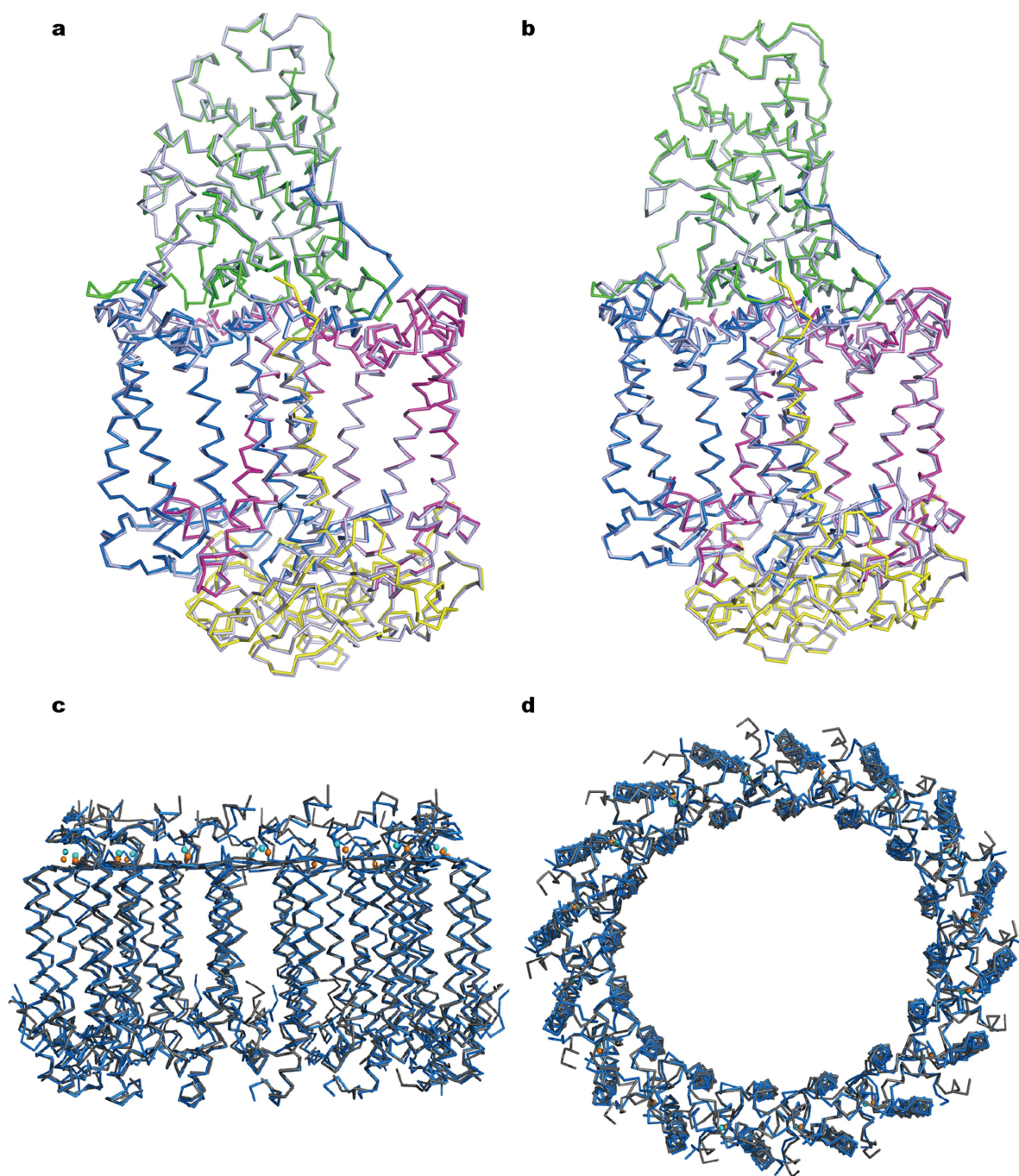
Extended Data Fig. 4 | Comparison of the protein structures between the previous and present structures. a, Superposition of the RC subunits between the previous 2.2 Å and the present 1.9 Å structures, with a side view from the membrane plane. **b**, Superposition of the RC subunits between the previous 3.0 Å and the present 1.9 Å structures, with a side view from the membrane plane. **c**, **d**, Superposition of the LH1 subunits between the previous 3.0 Å and present 1.9 Å structures, with a side view (**c**) and top view (**d**) relative to the membrane plane, respectively. In all panels, the present 1.9 Å structure is coloured, whereas the previous structures are depicted in grey.
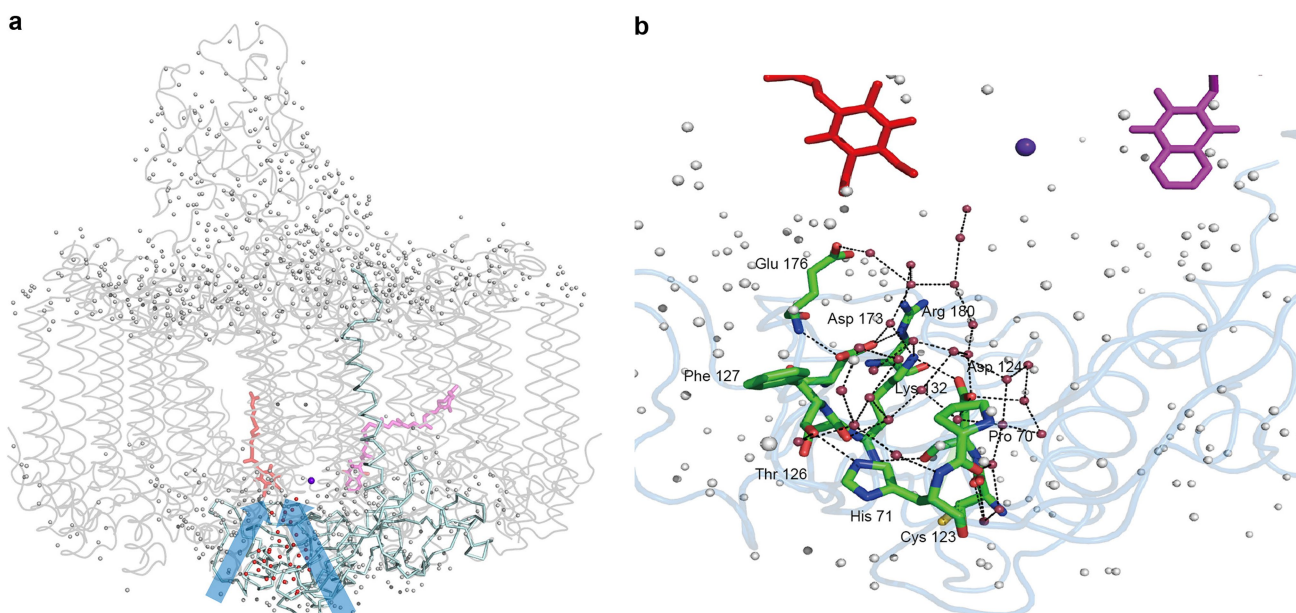
Extended Data Fig. 5 | Hydrogen-bonding networks for the protonation of Q$_B$. a, Two possible proton channels connecting Q$_B$ to the cytoplasmic surface. The thick arrow (coloured in blue) indicates the main channel formed within the H-subunit, which is enlarged in **b**, and the thin arrow indicates the second channel. **b**, The main hydrogen-bonding network indicated by the thick arrow in **a**, formed by a number of water molecules and the residues (green) from the H-subunit (pale cyan). Q$_A$ and Q$_B$ are depicted in violet and red, respectively, and the non-haem iron is depicted in deep purple. The hydrogen bonds are depicted as dashed lines. Water molecules participating in the hydrogen-bonding networks are depicted in orange, and those not participating are depicted in grey.
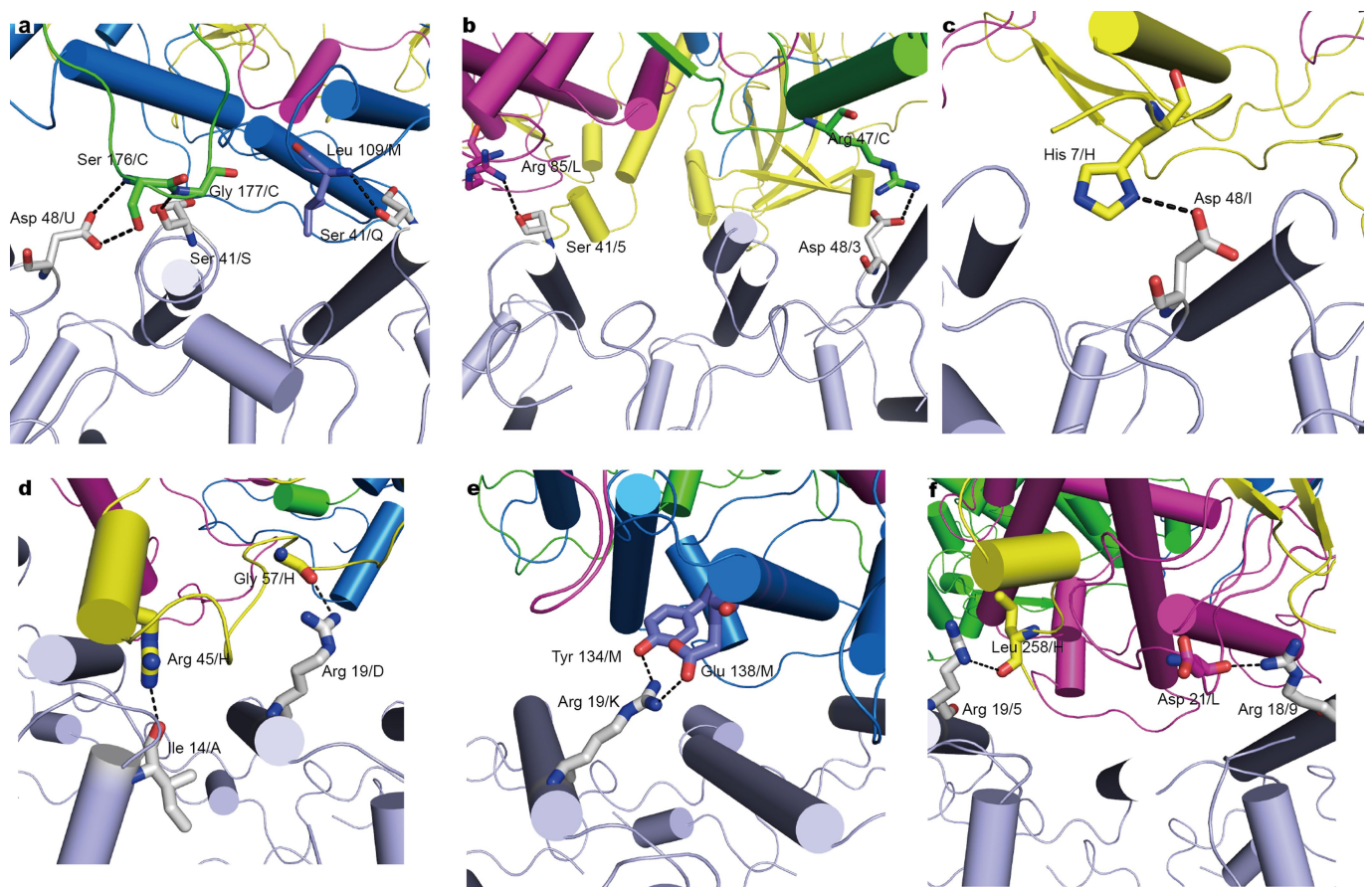
Extended Data Figure 6 | Protein–protein interactions between LH1 and RC. a–c, Interactions between the LH1 α-subunits and the RC subunits at the periplasmic side. Panels **a–c** correspond to boxed areas 1–3 in Fig. 4. **d–f,** Interactions between the LH1 α-subunits and RC subunits at the cytoplasmic side. Panels **d–f** correspond to boxed areas 4–6 in Fig. 4.

Extended Data Table 1 | Data collection and refinement statistics

	LH1-RC
Data collection	
Space group	C121
Cell dimensions	
a, b, c (Å)	145.23, 143.81, 210.28
α, β, γ (°)	90.00, 90.74, 90.00
Resolution (Å)	46.92-1.90 (1.968-1.900)[a]
R_{merge}	0.1035 (1.863)
$I / \sigma I$	9.47 (1.14)
Completeness (%)	99.95 (99.94)
Redundancy	9.2 (8.0)
Refinement	
Resolution (Å)	46.92-1.90 (1.968-1.900)[a]
No. reflections	338536 (33812)
R_{work} / R_{free}	0.1815 (0.3558)/0.2152 (0.3737)
No. atoms	
Protein	22003
Ligand/ion	5022
Water	956
B-factors	65.45
Protein	63.68
Ligand/ion	73.62
Water	63.26
R.m.s. deviations	
Bond lengths (Å)	0.018
Bond angles (°)	1.77

[a]Values in parentheses are for the highest-resolution shell. The table was prepared using PHENIX with all automatic default settings.

Extended Data Table 2 | Components of LH1–RC determined at 1.9 Å resolution

	Proteins	Cofactor	Numbers
RC	C, H, M, L	BChl *a*	4
		BPhe *a*	2
		Heme-Fe	4
		Spirilloxanthin	1
		Non-heme-Fe	1
		Mg	1
		MQ8	1
		UQ8	5
		CDL	9
		PG	10
		PEF	2
LH1	α (16) β (16)	BChl *a*	32
		Spirilloxanthin	16
		Ca	16
Total	36		104

Extended Data Table 3 | Average B-factors of the RC subunits and the α- and β-apoproteins of LH1

LH1 α-apoproteins	No. of atoms	Average B-factors
A	548	72.17
D	559	75.37
F	559	79.27
I	574	75.90
K	571	69.46
O	569	62.35
Q	571	56.56
S	582	60.01
U	580	65.40
W	565	67.74
Y	571	74.25
1	564	79.34
3	569	78.86
5	548	72.61
7	566	72.46
9	571	72.36
Average		70.83

LH1 β-apoproteins	No. of atoms	Average B-factors
B	417	80.85
E	392	84.21
G	417	89.55
J	417	82.48
N	417	74.38
P	417	64.49
R	411	60.12
T	426	64.16
V	417	73.47
X	411	77.82
Z	403	79.47
2	411	84.72
4	417	80.31
6	417	74.92
8	411	76.65
0	426	76.71
Average		76.47

RC subunits	No. of atoms	Average B-factors
C	2590	51.82
L	2466	44.45
M	2850	43.60
H	1976	60.59
Average		49.36

Extended Data Table 4 | Interactions between LH1 and RC

Periplasmic side	
LH1-α	RC
Protein-Protein	
Asp48/3	Arg47/C
Ser41/5	Arg85/L
Asp48/I	His7/H
Ser41/S	Gly177/C
Asp48/U	Ser176/C
Ser41/Q	Leu109/M
Protein-Lipid	
Asp48/1	PGV34/a
Asn45/1	
Asp48/5	PGV39/a
Ser41/A	PGV31/a
Asn45/D	
Leu40/D	
Asn45/K	PEF304/a
Leu40/K	

Cytoplasmic side	
LH1-α	RC
Protein-Protein	
Arg19/5	Leu258/H
Arg18/9	Asp21/L
Ile14/A	Arg45/H
Arg19/D	Gly57/H
Arg19/K	Tyr134/M
	Glu138/M
Protein-Lipid	
Arg19/9	PGV40/a
Arg18/A	
Arg19/A	CDL303/a
Ile14/A	CDL/30/a
Arg18/D	
Arg19/D	
Ser23/F	CDL25/a
Arg19/K	CDL27/a
Arg18/O	
Ser23/O	
Arg19/O	PEF12/a
Asp16/Q	CDL28/a
Arg19/S	CDL29/a
Arg19/U	
Arg18/Y	CDL24/a
Arg19/Y	CDL26/a

LH1-α	LH1-α
Asp43	Asp48
	Ser54
	Tyr55
	Gln56
Asn45	Gln56

LH1-α	LH1-β
Leu44	Arg43
Asn45	
Asp49	
Tyr55	Pro44

LH1-α	LH1-β
Tyr9	Asp11
Lys10	
Trp12	Thr7
Leu13	

LH1-β	LH1-β
Tyr42	Leu46

ARTICLE

Crystal structures of the gastric proton pump

Kazuhiro Abe[1,2,3]*, Katsumasa Irie[1,2], Hanayo Nakanishi[1,3], Hiroshi Suzuki[4] & Yoshinori Fujiyoshi[1,3,5]

The gastric proton pump—the H^+, K^+-ATPase—is a P-type ATPase responsible for acidifying the gastric juice down to pH 1. This corresponds to a million-fold proton gradient across the membrane of the parietal cell, the steepest known cation gradient of any mammalian tissue. The H^+, K^+-ATPase is an important target for drugs that treat gastric acid-related diseases. Here we present crystal structures of the H^+, K^+-ATPase in complex with two blockers, vonoprazan and SCH28080, in the luminal-open state, at 2.8 Å resolution. The drugs have partially overlapping but clearly distinct binding modes in the middle of a conduit running from the gastric lumen to the cation-binding site. The crystal structures suggest that the tight configuration at the cation-binding site lowers the pK_a value of Glu820 sufficiently to enable the release of a proton even into the pH 1 environment of the stomach.

The pH of stomach fluid decreases to around 1 in response to food intake. This highly acidic environment is generated by the gastric H^+, K^+-ATPase[1] and is indispensable for digestion, and is also an important barrier to pathogens invading via the oral route. However, excessive stomach acidification leads to ulcers, which—although not life-threatening—considerably impair the health of affected individuals[2]. Acid suppression in combination with antibiotics is the recognized treatment to eradicate *Helicobactor pylori*, a risk factor for gastric cancer[3]. Proton pump inhibitors, such as omeprazole, and a recently developed class of acid suppressants called K^+-competitive acid blockers (P-CABs), which includes vonoprazan, are commonly used to treat acid-related diseases[4]. One compound, SCH28080, was found to be hepatotoxic, and has therefore never been developed for clinical use. As a P-CAB prototype, however, SCH28080 has been used as a specific H^+, K^+-ATPase antagonist in vitro[5], and several related compounds are currently undergoing clinical trials[6]. Gastric H^+, K^+-ATPase continues to be a prominent target for drugs that treat excess stomach acid secretion.

As with other P-type ATPases, the cation transport performed by gastric H^+, K^+-ATPase is accomplished by cyclical conformational changes of the enzyme (abbreviated as E), generally described using an E1/E2 nomenclature based on the Post-Albers scheme[7] (Fig. 1). ATP-driven H^+ export into the gastric lumen and uptake of K^+ into the cytoplasm is electroneutral with a transport stoichiometry that is thought to vary from $2H^+/2K^+$ to $1H^+/1K^+$ per ATP as the luminal pH decreases[8]. A hallmark of the P-type ATPase family is the auto-phosphorylation of an invariant aspartate—Asp385, in H^+, K^+-ATPase—during the transport cycle to form a phosphoenzyme intermediate (E1P, E2P).

The H^+, K^+-ATPase comprises two subunits. Its catalytic α-subunit is highly homologous to those of related P2-type ATPases such as Na^+, K^+-ATPase[9] and sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA)[10], which share 65% and 35% sequence identity, respectively, with the α-subunit of H^+, K^+-ATPase. The α-subunit comprises 10 transmembrane helices (TM1–TM10), which contain the cation-binding sites, and large cytoplasmic domains—the nucleotide, phosphorylation, and actuator domains. In addition to the α-subunit, H^+, K^+-ATPase

and Na^+, K^+-ATPase require a type II membrane protein β-subunit for functional expression as an α–β-complex.

H^+, K^+-ATPase pumps H^+ from the neutral cytoplasm of the parietal cell (pH 7) to the acidic milieu of the stomach (down to pH 1; an approximately 10^6-fold H^+ gradient)[11]. Releasing H^+ into a pH 1 environment is an especially challenging task because the pK_a of free carboxyl groups is normally about 3–5. The molecular mechanism underlying H^+ transport into the stomach has long remained elusive. Here we describe the crystal structures of gastric H^+, K^+-ATPase in a luminal-open E2P conformation bound to vonoprazan or SCH28080, analysed at 2.8 Å resolution. These structures define the molecular interaction between P-CABs and H^+, K^+-ATPase, and reveal how H^+, K^+-ATPase expels H^+ into the stomach even at pH 1.

Overall structure

For the crystallization, we used pig gastric H^+, K^+-ATPase expressed in HEK293S cells using a baculovirus-mediated system[12,13]. To avoid excess glycosylation of the six *N*-linked glycosylation sites located on the extracellular part of the β-subunit, we used the GnT1⁻ strain, and endoglycosidase treatment was included during the purification steps (Extended Data Fig. 1). Crystals were obtained in the presence of detergent and phospholipid, giving type I crystals (Extended Data Fig. 1, Extended Data Table 1) as with most other crystallized P-type ATPases[10,14]. The asymmetric unit of the crystal comprises an α–β complex, and several phospholipids and detergent molecules could be identified. As crystals were grown in the presence of beryllium fluoride (BeF_3^-) and P-CABs (vonoprazan or SCH28080), the molecular conformation adopted an E2P state, to which P-CABs are preferentially bound[15] (Fig. 1a, see also Supplementary Video 1). The vonoprazan-bound (here termed (Von)E2BeF) and SCH28080-bound (here termed (SCH)E2BeF) structures were virtually the same (Extended Data Fig. 2, root mean square deviations (r.m.s.d.) = 0.79 Å). The overall structure of H^+, K^+-ATPase was very similar to the corresponding structures of SERCA Mg^{2+} E2BeF[16] (r.m.s.d. 2.3 Å) and ouabain-bound Na^+, K^+-ATPase E2P[17] (r.m.s.d. 1.3 Å) as defined by the relative orientations of the three cytoplasmic domains and the arrangement of the transmembrane helices. BeF_3^- mimics a bound

[1]Cellular and Structural Physiology Institute, Nagoya University, Nagoya, Japan. [2]Graduate School of Pharmaceutical Sciences, Nagoya University, Nagoya, Japan. [3]Japan Core Research for Evolutional Science and Technology, Japan Science and Technology Corporation, Nagoya, Japan. [4]Laboratory of Molecular Electron Microscopy, The Rockefeller University, New York, NY, USA. [5]CeSPIA Inc., Tokyo, Japan. *e-mail: kabe@cespi.nagoya-u.ac.jp
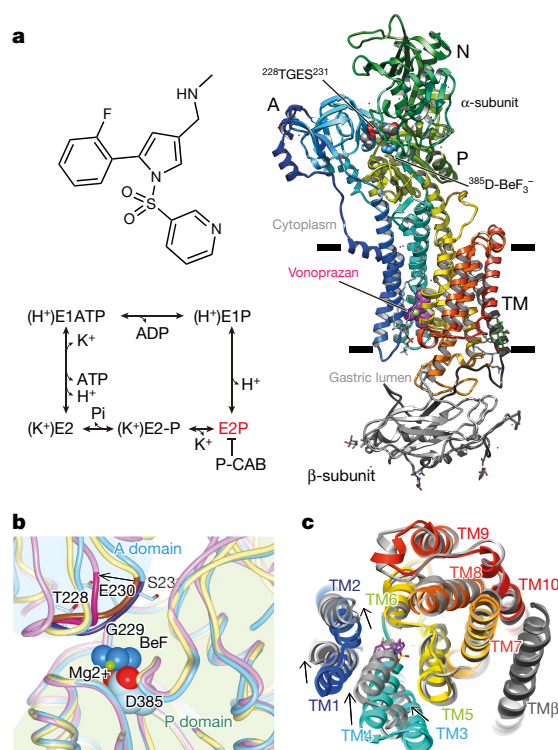
Fig. 1 | Crystal structure of gastric H⁺, K⁺-ATPase in complex with vonoprazan. a, Overall structure of the luminal-open E2P state of H⁺, K⁺-ATPase complexed with vonoprazan, (Von)E2BeF, in ribbon representation. Colour of the α-subunit gradually changes from the N terminus (blue) to the C terminus (red). Invariant Asp385, the phosphate analogue BeF₃⁻ (blue) in the phosphorylation domain and the TGES motif located at the edge of the actuator domain are highlighted as spheres. Colour of the β-subunit changes gradually from dark grey (N terminus in the cytoplasm) to light grey (C terminus in the gastric lumen). Two phospholipids (dioleoylphosphatidylcholine), two detergent molecules (octaethylene glycol monododecyl ether) and three N-linked N-acetylglucosamines are modelled in the structure (sticks). Approximate location of the membrane is indicated by black lines. Chemical structure of vonoprazan is provided in the upper left. Post-Albers type reaction scheme for H⁺, K⁺-ATPase is shown in the lower left. Inside and outside of the scheme represent cytoplasmic and luminal sides of the parietal cell, respectively. A, actuator domain; N, nucleotide domain; P, phosphorylation domain; TM, transmembrane domain. **b**, Close-up view of the phosphorylation site. BeF₃⁻ (blue) bound to Asp385 (light blue) and coordinating Mg²⁺ (green) are shown as spheres. H⁺, K⁺-ATPase (Von)E2BeF (light blue), SERCA E2BeF (RCSB Protein Data Bank (PDB) code: 3B9B, yellow)[16] and SERCA E2-AlF (PDB code: 2ZBG, pink)[18] are superimposed according to their phosphorylation domain structures. The TGES motif in each structure is indicated by dark colours (only amino acids of H⁺, K⁺-ATPase are shown in the stick representation). **c**, Comparison of the transmembrane helices (TM1–TM10 for the α-subunit, and TMβ for the β-subunit) between luminal-open (Von)E2BeF (coloured ribbons as in **a**) and the low-resolution model of luminal-closed, P-CAB-free, E2-AlF of H⁺, K⁺-ATPase (grey ribbons)[20]. A cross section of the luminal transmembrane region parallel to the membrane plane is shown, viewed from the luminal side of the membrane. Arrows in **b** and **c** indicate displacement of the TGES motif and indicated transmembrane helices associated with luminal gate closure (E2BeF to E2-AlF)[18].

phosphate and forms a covalent bond with the invariant Asp385 in the DKTG motif of the phosphorylation domain. Bound BeF₃⁻ is covered by a segment containing the TGES (Thr228-Gly229-Glu230-Ser231) motif from the actuator domain, which prevents spontaneous hydrolysis of the aspartylphosphate by the bulk water, in marked contrast to the E2-P transition state captured in SERCA with aluminium fluoride (AlF₄⁻)[18] (Fig. 1b). The position of the actuator domain results in tightening of the linker region connecting TM2 and the actuator domain, and elongation of an α-helical segment (Extended Data Fig. 3). Because
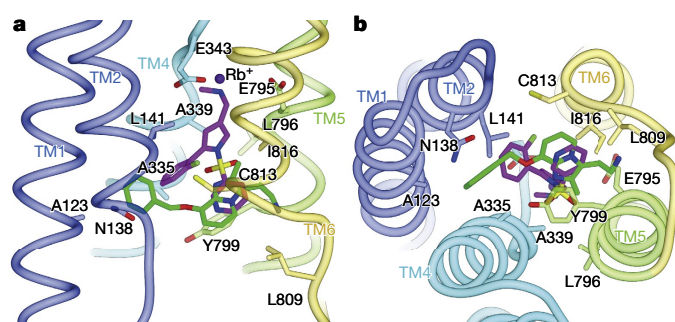

Fig. 2 | P-CAB-binding site. Bound SCH28080 is superimposed on the (Von)E2BeF structure to show their partially overlapping binding sites in the vicinity of Tyr799. The transmembrane helices are shown as ribbons with colouring as in Fig. 1a, except that TM5 is shown in green for clarity. Amino acids involved in the P-CAB coordination are shown as sticks. Structures are viewed from approximately parallel to the membrane plane (**a**) or perpendicular to the membrane plane from the luminal side (**b**).

of the extension, the TM1–TM2 bundle assumes an upright position towards the cytoplasmic side, which is coupled with the laterally open position of the TM3–TM4 bundle. As a consequence, the transmembrane helices have a luminal-open arrangement (Fig. 1c), in marked contrast to the P-CAB-free luminal-closed E2-P transition state of H⁺, K⁺-ATPase E2-AlF[19,20] determined by electron crystallography at 6.5 Å resolution (see Extended Data Fig. 3e for the schematics of the conformational change that accompanies luminal gating). The gate opening enables P-CABs to access their binding site from the luminal side (Extended Data Fig. 2c, d). These structural features clarify that the present H⁺, K⁺-ATPase structures, (Von)E2BeF and (SCH)E2BeF, adopt a luminal-open E2P state stabilized by a bound P-CAB[21].

Binding site for P–CAB

The electron density maps define the binding mode of the two P-CABs vonoprazan and SCH28080, and the residues coordinating them, in a luminal-facing conduit that extends to the cation-binding site (Fig. 2, Extended Data Figs. 4, 5 and Extended Data Table 2), which is apparently consistent with the K⁺-competitive inhibition of H⁺, K⁺-ATPase activity by blocking K⁺ entry to the cation-binding site. The binding sites of vonoprazan and SCH28080 were previously thought to overlap on the basis of their similar inhibitory actions[22]. Our structures show that they do indeed partially overlap but are also distinct, as detailed in Supplementary Information.

The gating latch

Extensive studies of P2-type ATPases[18,23,24] and electron crystallographic structures of H⁺, K⁺-ATPase[15] have revealed the conformational changes required to regulate the luminal gate (Extended Data Fig. 3e). Luminal gating—mostly operated by the TM1–TM2 and TM3–TM4 bundles—is allosterically regulated by the bound phosphate on the phosphorylation domain, and by the coordinating TGES loop on the actuator domain (Fig. 1b) that connects to TM1 and TM2. The key conformational change required for luminal gating is the vertical sliding movement of the TM1–TM2 bundle relative to TM3–TM4 (Extended Data Fig. 3). In the luminal-open P-CAB-bound E2P state of H⁺, K⁺-ATPase, the side chain of Ile119 in TM1 lies on top of the side chain of Met334 in the luminal portion of TM4, acting as a latch, and the TM1–TM2 bundle is held in an upright position towards the cytoplasmic side (Fig. 3a, b). If the positioning is indeed latch-like, the substitution of either Ile119 or Met334 with a smaller side-chain alanine would induce spontaneous slipping of the TM1–TM2 bundle towards the luminal side to close the luminal gate. The gate closure signal would in turn be transmitted to the actuator domain, triggering a change in the coordination between the TGES motif and the bound phosphate (Fig. 1b), and a bulk water molecule introduced subsequently
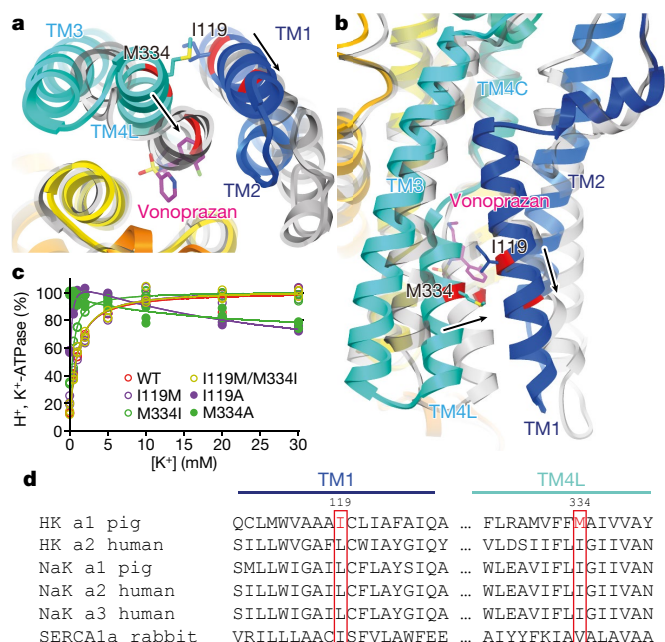
Fig. 3 | A gating latch. a, b, Comparison of the luminal-open (Von) E2BeF (coloured ribbons) and low-resolution E2-AlF model[20] (grey ribbons, luminal-closed E2-P transition state) of H^+, K^+-ATPase. Ile119 and Met334 are shown as stick representations. Arrows indicate the displacement of Ile119 and Met334 C_α positions (indicated in red on TM1 and TM4 helices) from the luminal-open to the luminal-closed forms. Structures are viewed from the luminal side (**a**) and approximately parallel to the membrane plane (**b**). **c,** K^+-dependent ATPase activities of the wild-type enzyme and indicated mutants. Data plotted were corrected for background values in the absence of K^+ and the presence of 10 μM SCH28080, and normalized to their maximum velocity set to 100%. The value in the absence of KCl therefore indicates the H^+-ATPase activity that corresponds to the rate of the spontaneous E2P dephosphorylation. Individual data from triplicated points at eight K^+ concentrations were plotted. Representative results from more than three independent measurements for each of the mutants are shown in the figure. **d,** Sequence alignment of pig gastric H^+, K^+-ATPase (α1) with other P2-type ATPases. The positions for Ile119 and Met334 in pig gastric H^+, K^+-ATPase (highlighted in red) are framed with red boxes. TM4L, luminal position in TM4; TM4C, cytoplasmic portion in TM4; WT, wild type.

into the phosphorylation centre would induce dephosphorylation. In the native transport cycle, this sequence of actions is induced by the binding of the counter-transporting K^+ to the cation-binding site as a rate-limiting step. This is why H^+, K^+-ATPase activity is accelerated in a K^+-dependent manner, as seen in the wild-type enzyme (Fig. 3c). By contrast, the ATPase activities of Ile119Ala and Met334Ala mutants are nearly K^+-independent and constitutively active (Fig. 3c and Extended Data Table 3). These amino acids are located a long way from the cation-binding site, and it is therefore unlikely to be the case that these mutations affect cation-binding site properties. On the other hand, the Ile119Met and Met334Ile mutants as well as the Ile119Met/Met334Ile double mutants—all of which have large hydrophobic side chains—exhibit normal K^+-dependent ATPase activity, as seen in the wild-type enzyme. We therefore conclude that in the luminal-open state, these hydrophobic amino acids act as a latch that prevents a sliding movement of the TM1–TM2 bundle with respect to the luminal portion of TM4, and that they are therefore important for tight coupling between K^+ binding and dephosphorylation (Extended Data Fig. 3e). A comparison of the amino acid sequences of P2-type ATPases reveals that the amino acid residues at the corresponding positions in TM1 and the TM4 luminal portion are all bulky and hydrophobic (Fig. 3d), which suggests that the mechanism of luminal gate regulation described here is likely to be conserved among P2-type ATPases.
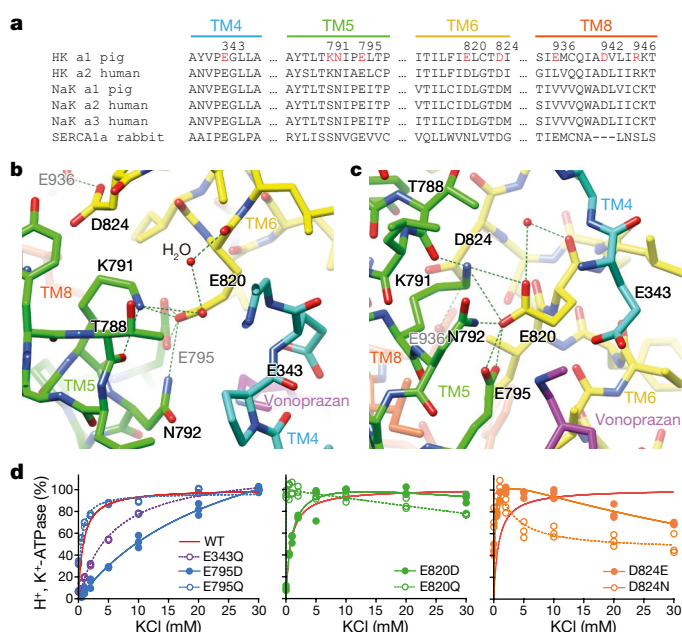

Fig. 4 | Cation-binding site in the luminal-open E2P state. a, Sequence alignment of the indicated transmembrane helices among related P2-type ATPases. The amino acids explicitly discussed here are highlighted in red. **b, c,** Close-up of the cation-binding site in H^+, K^+-ATPase (Von) E2BeF in stick representation, viewed approximately perpendicular to the membrane from the cytoplasmic side (**b**) and parallel to the membrane from the TM4 side (**c**). Dotted lines are shown between residues with ≤ 3.5 Å between neighbouring atoms, presumably making hydrogen bonds or an electrostatic interaction (Lys791–Glu820). A water molecule (red) is also indicated. **d,** K^+-dependence of SCH28080-sensitive ATPase activity of the indicated mutant enzymes determined as in Fig. 3c. Glu343Asp exhibited no detectable ATPase activity. Individual data from triplicated points at each of the indicated K^+ concentrations were plotted, and representative results from more than three independent measurements for each mutant are shown in the figure.

Mechanism of proton extrusion

According to the Post-Albers scheme (Fig. 1a), the luminal-open E2P state is an intermediate state occurring just after proton release, and poised for subsequent K^+ binding. In fact, the cation-binding site is exposed to the luminal bulk medium when bound P-CAB is removed from our structures (Extended Data Fig. 2b, c). According to previous H^+ transport measurements using inside-out vesicles taken from pig stomach[8], two protons are released in exchange for two K^+ ions at neutral pH. However, in theory, under acidic conditions only a single proton can be transported, and only a single K^+ counter-transported, per hydrolysed ATP[8]. This variable transport stoichiometry hypothesis for H^+, K^+-ATPase[8,20] therefore requires two distinct proton-binding sites with different pK_a values, but details of the molecular mechanisms have remained elusive. Previous mutagenesis studies have demonstrated that Na^+, K^+-ATPase[25] and H^+, K^+-ATPase[26] transport Na^+ and H^+, respectively, generally by using the same conserved carboxylic acids of their respective cation-binding sites in TM4 and TM6, except for a lysine residue in TM5 of H^+, K^+-ATPase (Fig. 4a). This lysine (Lys791 in the pig sequence), which is invariant among gastric H^+, K^+-ATPases, is a serine in Na^+, K^+-ATPase and SERCA, and is predicted to be important for H^+, K^+-ATPase properties, such as proton transport[27], net electroneutral cation transport[28] and inherent E2 preference[26].

In our structures, the carboxyl residue of Glu820 is at the centre of the cation-binding site (Fig. 4b, c and Supplementary Video 2). This glutamate residue is surrounded by other polar amino acids, including Asn792, Glu795 and Lys791. The juxtaposition of the two glutamates Glu795 and Glu820 (2.5 Å between their closest oxygens) indicates that one of these acidic residues is protonated. Because the charge-neutralized Glu795Gln mutant exhibits an ATPase activity profile comparable
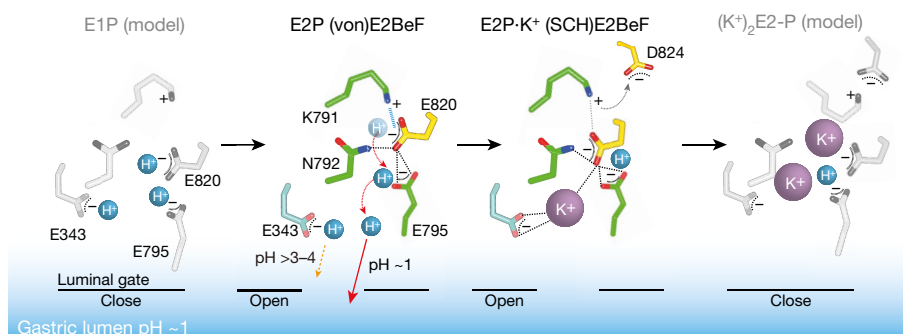
Fig. 5 | A model for the proton extrusion into the acidic solution by the gastric H^+, K^+-ATPase. In the H^+-occluded E1P state (left), all three glutamates in the cation-binding site are protonated; otherwise, H^+ would be incorporated into the cation-binding site owing to its high concentration in the stomach when the luminal gate opens. In the luminal-open E2P state, the proton-binding affinity of the Glu820 carboxyl is strongly reduced because of its juxtaposition with Glu795, a hydrogen bound to Asn792 and a salt bridge with Lys791. As a consequence, a single H^+ is expelled into the luminal acidic solution, presumably via Glu795, which is exposed to the surface of the luminal cavity in the structure (red arrows). The Glu343 releases H^+ only when the luminal solution is neutral to weakly acidic according to its own pK_a value (orange arrow). After H^+

release, K^+ is incorporated into the cation-binding site in the E2P·K^+ form. Coordination of the K^+ ion by Glu820 may release the salt bridge between Lys791 and Glu820 (grey dashed line), and Lys791 forms a new salt bridge with Asp824 (grey arrow). This sequence of actions triggers a conformational change of the whole enzyme to drive the transport cycle, and counter-transporting K^+ ions are occluded in the luminal-closed $(K^+)_2$E2-P state (right). This mechanism is supported functionally by the constitutively active phenotype of charge-neutralized mutants of Glu820Gln and Asp824Asn[32] (Fig. 4d). See Supplementary Information, Extended Data Figs. 6, 7, and Supplementary Video 3 for details of K^+-binding.

to the wild-type enzyme and Glu820Gln does not (Fig. 4d), Glu795 rather than Glu820 is likely to be protonated. Therefore, these two glutamate residues interact through a hydrogen bond. Glu820 also receives hydrogen bonds from Asn792 (distance of 3.0 Å) and a water molecule (3.5 Å). In addition to this hydrogen bond network around Glu820, the ε-amino group of Lys791 interacts intimately with the carboxylate of Glu820 (3.1 and 3.2 Å from Oδ1 and Oδ2, respectively), most probably forming a salt bridge in the crystal structure as suggested from functional studies[26,29]. The Glu820 carboxyl is thus situated in an unusual environment with extensive polar interactions that could lower its pK_a value. A reduction of pK_a values in juxtaposed carboxyl groups of two adjacent acidic residues occurs in the catalytic centres of many other enzymes; for example, two aspartate residues 2.5 Å apart in the catalytic centre of pepsin[30] were estimated to have pK_a values of 1.2 and 4.7[31]. Several aspartate residues located on the surface of pepsin and receiving multiple hydrogen bonds from the surroundings and/or coordinated by basic amino acids exhibit unusually low pK_a values, displaying a negative charge even in the highly acidic environment of the stomach[30]. Therefore, H^+, K^+-ATPase Glu820 is a strong candidate for one of the proton release sites, presumably through the luminally exposed Glu795 according to the Grotthuss mechanism (Fig. 5).

Another glutamate, Glu343, in TM4 is also highly conserved among P2-type ATPases and is important for cation transport[32]. In our structures, Glu343 is located at some distance from Lys791 (7.4 Å) and other glutamate residues (5.7 Å and 4.7 Å for Glu795 and Glu820, respectively), and may therefore release an H^+ only when the lumen is neutral to weakly acidic (Fig. 5). These glutamate residues (Glu343, Glu795 and Glu820) in the cation-binding site are invariant for the gastric H^+, K^+-ATPase α1 isoform. The Glu820 residue in H^+, K^+-ATPase α1, however, corresponds to a shorter aspartate in Na^+, K^+-ATPase and non-gastric H^+, K^+-ATPase α2, and the latter pump transports both H^+ and Na^+ (Fig. 4a). The longer side chain of glutamate in the gastric H^+, K^+-ATPase (Glu820) is likely to be better suited for the tight hydrogen bond network that reduces its proton affinity, and may also underlie the H^+ specificity of the E1 state.

Plant H^+-ATPase, a similar H^+-transporting pump, is structurally and functionally well characterized[33]. A crystal structure of AHA2 from *Arabidopsis thaliana* has revealed that two key residues for the H^+-transport—Arg655 and Asp684, which correspond to Glu795 and Asp824, respectively, in our structures—are located in the cation-binding site[34]. It has previously been proposed[35] that these residues form a salt bridge as a consequence of the conformation change to the E2P state

required to release H^+, although the structure determined in this previous study is in the E1 state. The α2-isoform of brine shrimp Na^+, K^+-ATPase has two lysines, Lys324 and Lys776, instead of Tyr340 and Asn792 in its H^+, K^+-ATPase, and the expression level of this isoform rises considerably when the salt concentration in the living environment increases[36]. These lysines in this brine shrimp Na^+ pump may also play an important role in Na^+ extrusion against a very steep Na^+ gradient. The utilization of a basic amino acid to facilitate extrusion may be a general cation transport mechanism.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0003-8.

1. Ganser, A. L. & Forte, J. G. K^+-stimulated ATPase in purified microsomes of bullfrog oxyntic cells. *Biochim. Bioshys. Acta* **307**, 169–180 (1973).
2. Sachs, G. et al. The gastric H,K ATPase as a drug target: past, present, and future. *J. Clin. Gastroenterol.* **41**, S226–S242 (2007).
3. Sachs, G., Meyer-Rosberg, K., Scott, D. R. & Melchers, K. Acid, protons and *Helicobacter pylori*. *Yale J. Biol. Med.* **69**, 301–316 (1996).
4. Otake, K. et al. Characteristics of the novel potassium-competitive acid blocker vonoprazan fumarate (TAK-438). *Adv. Ther.* **33**, 1140–1157 (2016).
5. Kaminski, J. J., Wallmark, B., Briving, C. & Andersson, B. M. Antiulcer agents. 5. Inhibition of gastric H^+/K^+-ATPase by substituted imidazo[1,2-a]pyridines and related analogues and its implication in modeling the high affinity potassium ion binding site of the gastric proton pump enzyme. *J. Med. Chem.* **34**, 533–541 (1991).
6. Shin, J. M. & Sachs, G. Pharmacology of proton pump inhibitors. *Curr. Gastroenterol. Rep.* **10**, 528–534 (2008).
7. Rabon, E. C. & Reuben, M. A. The mechanism and structure of the gastric H,K-ATPase. *Annu. Rev. Physiol.* **52**, 321–344 (1990).
8. Rabon, E. C., McFall, T. L. & Sachs, G. The gastric [H,K]ATPase:H^+/ATP stoichiometry. *J. Biol. Chem.* **257**, 6296–6299 (1982).
9. Morth, J. P. et al. Crystal structure of the sodium–potassium pump. *Nature* **450**, 1043–1049 (2007).
10. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**, 647–655 (2000).
11. Wolosin, J. M. Ion transport studies with H^+-K^+-ATPase-rich vesicles: implications for HCl secretion and parietal cell physiology. *Am. J. Physiol. Gastrointest. Liver Physiol.* **248**, G595–G607 (1985).
12. Dukkipati, A., Park, H. H., Waghray, D., Fischer, S. & Garcia, K. C. BacMam system for high-level expression of recombinant soluble and membrane glycoproteins for structural studies. *Protein Expr. Purif.* **62**, 160–170 (2008).

13. Goehring, A. et al. Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat. Protocols* **9**, 2574–2585 (2014).
14. Bublitz, M., Poulsen, H., Morth, J. P. & Nissen, P. In and out of the cation pumps: P-type ATPase structure revisited. *Curr. Opin. Struct. Biol.* **20**, 431–439 (2010).
15. Abe, K., Tani, K. & Fujiyoshi, Y. Systematic comparison of molecular conformations of H$^+$,K$^+$-ATPase reveals an important contribution of the A-M2 linker for the luminal gating. *J. Biol. Chem.* **289**, 30590–30601 (2014).
16. Olesen, C. et al. The structural basis of calcium transport by the calcium pump. *Nature* **450**, 1036–1042 (2007).
17. Laursen, M., Yatime, L., Nissen, P. & Fedosova, N. U. Crystal structure of the high-affinity Na$^+$K$^+$-ATPase–ouabain complex with Mg^{2+} bound in the cation binding site. *Proc. Natl Acad. Sci. USA* **110**, 10958–10963 (2013).
18. Toyoshima, C., Norimatsu, Y., Iwasawa, S., Tsuda, T. & Ogawa, H. How processing of aspartylphosphate is coupled to lumenal gating of the ion pathway in the calcium pump. *Proc. Natl Acad. Sci. USA* **104**, 19831–19836 (2007).
19. Abe, K., Tani, K., Nishizawa, T. & Fujiyoshi, Y. Inter-subunit interaction of gastric H$^+$,K$^+$-ATPase prevents reverse reaction of the transport cycle. *EMBO J.* **28**, 1637–1643 (2009).
20. Abe, K., Tani, K., Friedrich, T. & Fujiyoshi, Y. Cryo-EM structure of gastric H$^+$,K$^+$-ATPase with a single occupied cation-binding site. *Proc. Natl Acad. Sci. USA* **109**, 18401–18406 (2012).
21. Abe, K., Tani, K. & Fujiyoshi, Y. Conformational rearrangement of gastric H$^+$,K$^+$-ATPase induced by an acid suppressant. *Nat. Commun.* **2**, 155 (2011).
22. Scott, D. R., Munson, K. B., Marcus, E. A., Lambrecht, N. W. G. & Sachs, G. The binding selectivity of vonoprazan (TAK-438) to the gastric H$^+$, K$^+$-ATPase. *Aliment. Pharmacol. Ther.* **42**, 1315–1326 (2015).
23. Danko, S., Yamasaki, K., Daiho, T. & Suzuki, H. Distinct natures of beryllium fluoride-bound, aluminum fluoride-bound, and magnesium fluoride-bound stable analogues of an ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+}-ATPase: changes in catalytic and transport sites during phosphoenzyme hydrolysis. *J. Biol. Chem.* **279**, 14991–14998 (2004).
24. Danko, S., Yamasaki, K., Daiho, T. & Suzuki, H. Membrane perturbation of ADP-insensitive phosphoenzyme of Ca^{2+}-ATPase modifies gathering of transmembrane helix M2 with cytoplasmic domains and luminal gating. *Sci. Rep.* **7**, 41172 (2017).
25. Jorgensen, P. L., Håkansson, K. O. & Karlish, S. J. D. Structure and mechanism of Na,K-ATPase: functional sites and their interactions. *Annu. Rev. Physiol.* **65**, 817–849 (2003).
26. Koenderink, J. B., Swarts, H. G. P., Willems, P. H. G. M., Krieger, E. & De Pont, J. J. H. H. M. A conformation-specific interhelical salt bridge in the K$^+$ binding site of gastric H,K-ATPase. *J. Biol. Chem.* **279**, 16417–16424 (2004).
27. Munson, K., Garcia, R. & Sachs, G. Inhibitor and ion binding sites on the gastric H,K-ATPase. *Biochemistry* **44**, 5267–5284 (2005).
28. Burnay, M., Crambert, G., Kharoubi-Hess, S., Geering, K. & Horisberger, J. D. Electrogenicity of Na,K- and H,K-ATPase activity and presence of a positively charged amino acid in the fifth transmembrane segment. *J. Biol. Chem.* **278**, 19237–19244 (2003).
29. Dürr, K. L., Seuffert, I. & Friedrich, T. Deceleration of the E1P–E2P transition and ion transport by mutation of potentially salt bridge-forming residues Lys-791 and Glu-820 in gastric H$^+$/K$^+$-ATPase. *J. Biol. Chem.* **285**, 39366–39379 (2010).
30. Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S. & James, M. N. G. Molecular and crystal structures of monoclinic porcine pepsin refined at 1.8 Å resolution. *J. Mol. Biol.* **214**, 143–170 (1990).
31. Clement, G. E., Rooney, J., Zakheim, D. & Eastman, J. The pH dependence of the dephosphorylated pepsin-catalyzed hydrolysis of N-acetyl-ʟ-phenylalanyl-ʟ-tyrosine methyl ester. *J. Am. Chem. Soc.* **92**, 186–189 (1970).
32. Asano, S., Furumoto, R., Tega, Y., Matsuda, S. & Takeguchi, N. Mutational analysis of the putative K$^+$-binding site on the fourth transmembrane segment of the gastric H$^+$,K$^+$-ATPase. *J. Biochem.* **127**, 993–1000 (2000).
33. Palmgren, M. G., Buch-Pedersen, M. J. & Møller, A. L. Mechanism of proton pumping by plant plasma membrane H$^+$-ATPase: role of residues in transmembrane segments 5 and 6. *Ann. NY Acad. Sci.* **986**, 188–197 (2003).
34. Pedersen, B. P., Buch-Pedersen, M. J., Morth, J. P., Palmgren, M. G. & Nissen, P. Crystal structure of the plasma membrane proton pump. *Nature* **450**, 1111–1114 (2007).
35. Focht, D., Croll, T. I., Pedersen, B. P. & Nissen, P. Improved model of proton pump crystal structure obtained by interactive molecular dynamics flexible fitting expands the mechanistic model for proton translocation in P-type ATPases. *Front. Physiol.* **8**, 202 (2017).
36. Jorgensen, P. L. & Amat, F. Regulation and function of lysine-substituted Na,K pumps in salt adaptation of *Artemia franciscana*. *J. Membr. Biol.* **221**, 39–49 (2008).

Author contributions K.A. and Y.F. designed the study. K.A. and H.S. expressed the proteins. K.A. and H.N. purified and crystallized the proteins. K.A. performed the biochemical analysis. K.A., H.N. and K.I. corrected the X-ray diffraction data. K.A. and K.I. analysed the structures. All authors interpreted the structure and wrote the manuscript.

Competing interests Y.F. is a director of CeSPIA Inc.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0003-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0003-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. The plasmids encoding the cDNA of pig gastric H⁺, K⁺-ATPase α- and β-subunits were provided by T. Imagawa. The Flag epitope tag (DYKDDDDK), hexa-histidine tag and the enhanced green fluorescent protein (EGFP), followed by a tobacco etch virus (TEV) protease recognition sequence, were attached to the amino-terminal of the Met48 of the α-subunit, and cloned into a custom-made vector based on a previous report[12]. The pig gastric H⁺, K⁺-ATPase β-subunit (wild type) was also cloned independently. The α–β-complex of H⁺, K⁺-ATPase was successfully expressed in the plasma membrane using baculovirus-mediated transduction of mammalian HEK293S GnT1⁻ cells, as previously described[13]. The collected cells were broken up using a high-pressure emulsifier (Avestin) in the presence of protease inhibitor cocktail (Roche), and membrane fractions were collected (200,000g for 1 h) after removing the cell debris (800g for 10 min). Membrane fractions were solubilized with 1% octaethylene glycol monododecyl ether ($C_{12}E_8$, Nikko Chemical) in the presence of 40 mM MES/Tris (pH 6.5), 20 mM Mg(CH_3COO)$_2$, 10% glycerol, 50 mM NaCl, 1 mM $BeSO_4$, 3 mM NaF, 1 mM ADP, 5 mM dithiothreitol, and 0.1 mM vonoprazan or SCH28080 on ice for 20 min. After removing the insoluble materials by ultracentrifugation (200,000g for 1 h), the supernatant was mixed with anti-Flag M2 affinity resin (Sigma Aldrich) for 2 h at 4 °C. The resin was washed with 20 column volumes of buffer consisting of 20 mM MES/Tris (pH 6.5), 5% glycerol, 2 mM MgCl₂, 50 mM NaCl and 0.03% $C_{12}E_8$. Flag-EGFP-tagged H⁺, K⁺-ATPase was eluted with 0.2 mg/ml Flag peptide (Sigma Aldrich) in the presence of 10 μM vonoprazan or SCH28080. Eluted fractions were incubated with TEV protease and MBP-fusion endoglycosidase (New England Biolabs) at 4 °C overnight. Digested peptide fragments containing EGFP and MBP-fusion endoglycosidase were removed by passing the fractions through a Ni-NTA resin (Qiagen) and amylose resin (New England Biolabs), respectively. Flow-through fractions were concentrated and subjected to a size-exclusion column chromatograph using a Superose6 Increase column (GE Healthcare), equilibrated in buffer comprising 10 mM MES/Tris (pH 6.5), 1% glycerol, 100 mM NaCl, 1 mM MgCl₂ and 0.03% $C_{12}E_8$. Peak fractions were collected and concentrated to 10 mg/ml. The concentrated H⁺, K⁺-ATPase samples were mixed with 0.5 mM $BeSO_4$, 1.5 mM NaF and 0.1 mM vonoprazan or SCH28080, and then added to the glass tubes in which a layer of dried dioleoyl phosphatidylcholine had formed, in a lipid-to-protein ratio of 0.3–0.5, and incubated overnight at 4 °C in a shaker mixer operated at 120 r.p.m.[37]. After removing the insoluble materials by ultracentrifugation, lipidated samples were used for the crystallization. Note that the effect of deglycosylation on the ATPase activity was negligible, as evaluated by K⁺ and P-CAB affinities compared with those of wild type without endoglycosidase treatment, as well as the native enzyme purified from pig stomach.

Crystallization. Crystals were obtained by vapour diffusion at 20 °C. For the vonoprazan-bound form, a 5-mg/ml purified, lipidated protein sample was mixed with reservoir solution containing 10% glycerol, 20% PEG2000MME, 0.4 M CH_3COONa, 3% methylpentanediol and 5 mM β-mercaptoethanol. For the SCH28080-bound form, reservoir solution containing 10% glycerol, 20% PEG2000MME, 0.2 M RbCl, 5% *tert*-butanol and 5 mM β-mercaptoethanol was used. Vonoprazan crystals were grown to 400 × 100 × 40 μm in 2 weeks, and SCH28080 crystals were grown to 400 × 200 × 200 μm in 3 weeks. Crystals were flash frozen in liquid nitrogen.

Structural determination and analysis. Diffraction data were collected at the SPring-8 beamline BL32XU and BL41XU, and processed using XDS. Structure factors were subjected to anisotropy correction using the UCLA MBI Diffraction Anisotropy server[38] (http://services.mbi.ucla.edu/anisoscale/). The vonoprazan-bound structure was determined by molecular replacement with PHASER, using the homology model of BYK99-bound H⁺, K⁺-ATPase (PDB code: 5Y0B) based on the electron crystallographic structure as a search model. Coot[39] was used for cycles of iterative model building and Refmac5 and Phenix[40] were used for refinement. The final crystallographic model of vonoprazan-bound H⁺, K⁺-ATPase at 2.80 Å resolution, refined to R_{work} and R_{free} of 0.237 and 0.288, was deposited in the PDB with accession code 5YLU. For the determination of the SCH28080-bound structure, the vonoprazan-bound form was used as the starting model for molecular replacement and the final crystallographic model at 2.80 Å resolution, refined to R_{work} and R_{free} of 0.240 and 0.292, was deposited in the PDB with accession code 5YLV. Rubidium ions were identified in anomalous difference Fourier maps calculated using data collected at a wavelengths of 0.8147 Å. The vonoprazan-bound and SCH28080-bound models contained 93.0, 6.8 and 0.2% and 91.3, 8.1 and 0.6% in the favoured, allowed and outlier regions of the Ramachandran plot, respectively.

Activity assay using recombinant proteins. To measure the ATPase activity, Flag-EGFP tag connected by the TEV cleavage site to the N-terminal tail of the wild-type α-subunit was used to monitor its expression by fluorescence size-exclusion column chromatography[41]. The wild-type or mutant α-subunit was co-expressed with the wild-type β-subunit using the BacMam system as described above, and broken membrane fractions were collected. H⁺, K⁺-ATPase activity was measured as previously described[42]. In brief, permeabilized membrane fractions (wild type or mutant) were suspended in buffer comprising 40 mM PIPES/Tris (pH 7.0), 2 mM MgCl₂, 2 mM ATP and 0–30 mM KCl in the presence of three different concentrations of vonoprazan or SCH28080, or their absence, in the 96-well plates. Reactions were initiated by incubating the fractions at 37 °C using a thermal cycler, and maintained for 1 to 5 h depending on their activity. Reactions were terminated, and the amount of released inorganic phosphate was determined colourimetrically using a microplate reader (TECAN). The inhibition constant (K_i) and K⁺-affinity (K_m) were determined as previously described[42]. Note that the Flag-EGFP tag and the N-terminal 47 amino acids (not present in the crystallized sample) had negligible effects on ATPase activity, as evaluated by their K⁺ and P-CAB affinities compared with those of the tag-free wild-type enzyme as well as the native enzyme purified from pig stomach.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Atomic coordination and structure factors for the structures reported in this work were deposited in the RCSB Protein Data Bank under accession numbers 5YLU (vonoprazan-bound) and 5YLV (SCH28080-bound). All other data that support the findings of this study are available from the corresponding author upon reasonable request.

37. Gourdon, P. et al. HiLiDe—systematic approach to membrane protein crystallization in lipid and detergent. *Cryst. Growth Des.* **11**, 2098–2106 (2011).
38. Strong, M. et al. Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103**, 8060–8065 (2006).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
40. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
41. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).
42. Abe, K. et al. The cryo-EM structure of gastric H⁺,K⁺-ATPase with bound BYK99, a high-affinity member of K⁺-competitive, imidazo[1,2-a]pyridine inhibitors. *Sci. Rep.* **7**, 6632 (2017).
43. Kanai, R., Ogawa, H., Vilsen, B., Cornelius, F. & Toyoshima, C. Crystal structure of a Na⁺-bound Na⁺K⁺-ATPase preceding the E1P state. *Nature* **502**, 201–206 (2013).
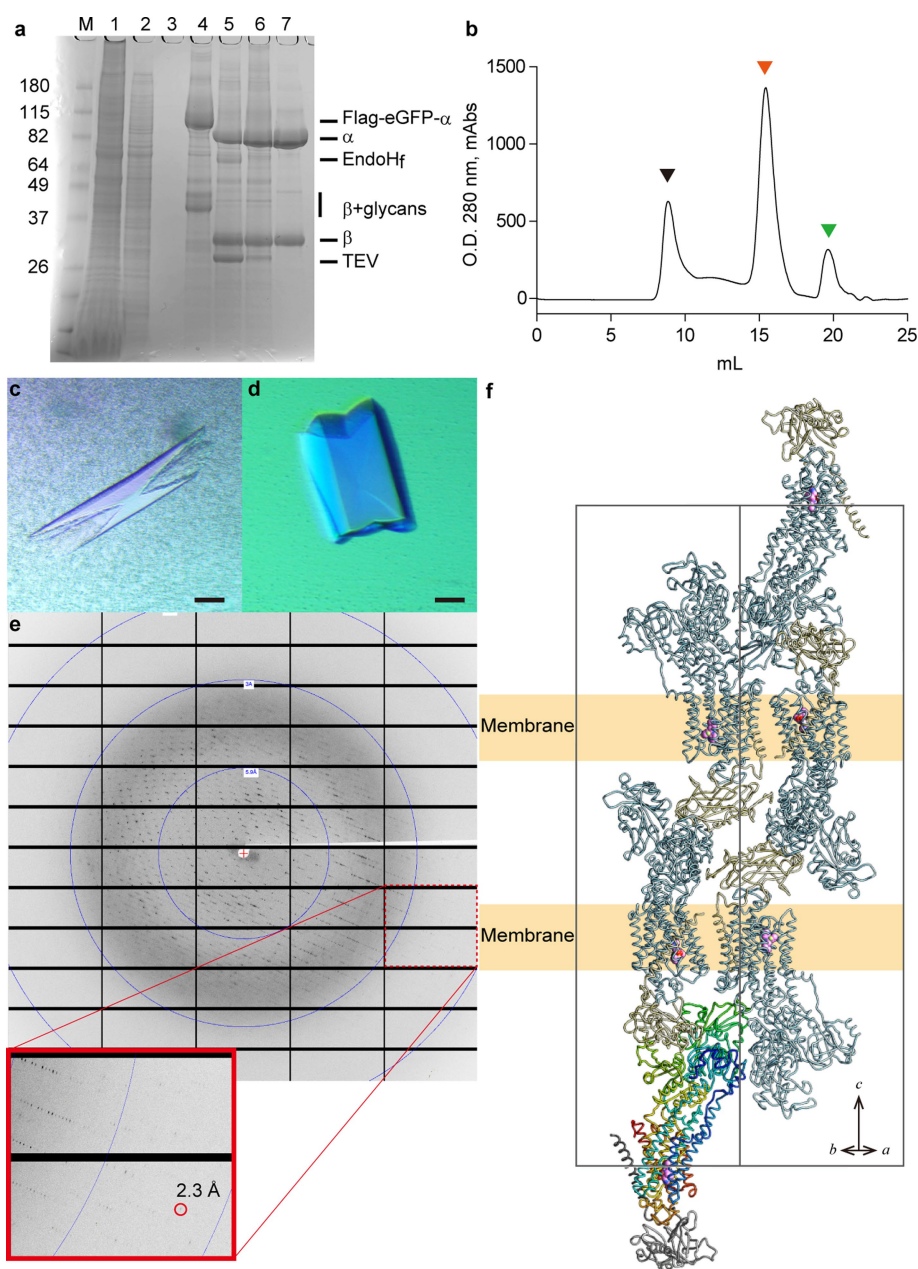
**Extended Data Fig. 1 | Crystallization of gastric H⁺, K⁺-ATPase.**
a, Purification of H^+, K^+-ATPase expressed in HEK293 cells. Lane 1: solubilized membrane fraction, lane 2: pass through of Flag resin, lane 3: wash fraction, lane 4: elution by Flag peptide, lane 5: TEV protease- and endoglycosidase-treated sample, lane 6: pass-through fraction of Ni-NTA and amylose resin, lane 7: concentrated peak fractions by size-exclusion chromatography. **b**, The elution profile of affinity-purified H^+, K^+-ATPase by Superose6 Increase 10/300. Black, red and green arrowheads indicate elution volume of aggregation, α–β-complex of H^+, K^+-ATPase and cleaved EGFP, respectively. Purification was well reproduced, and representative results are shown in the figure. **c**, **d**, Crystals of H^+, K^+-ATPase in the presence of vonoprazan (**c**) or SCH28080 and Rb^+ (**d**).

Scale bars, 100 μm. **e**, X-ray diffraction of the vonoprazan-bound crystal. Enlarged image shows diffraction spots of up to 2.3 Å in the direction of the $c*$ axis, although the crystal shows anisotropic diffractions. Most crystals showed diffraction spots of up to 2.8 Å in similar crystallization conditions and a few crystals showed diffraction spots better than 2.3 Å, as shown in the figure. **f**, Crystal packing. An asymmetric unit (molecule in the lower left, depicted as in Fig. 1d) contains one α–β-complex of H^+, K^+-ATPase (α-subunit, light blue; β-subunit, wheat; vonoprazan, magenta), packed with $P3_121$ symmetry. A unit cell and approximate location of the membrane planes are provided as grey and yellow boxes, respectively.
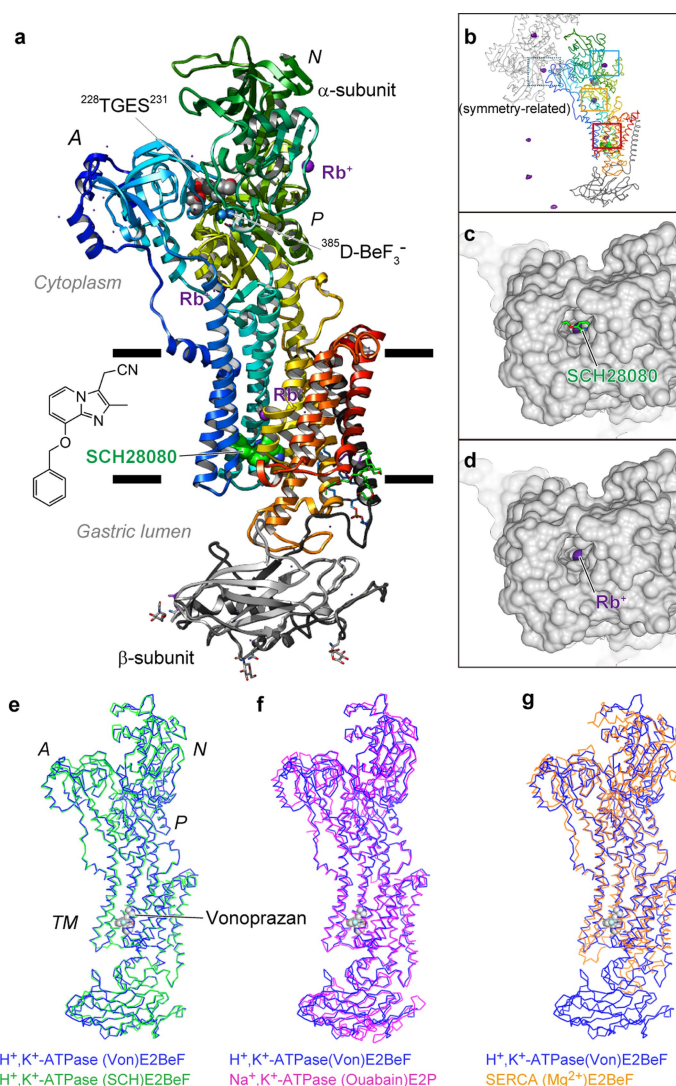
Extended Data Fig. 2 | Crystal structure of gastric H⁺, K⁺-ATPase bound to SCH28080. a, Overall structure of the luminal-open E2P state of H⁺, K⁺-ATPase complexed with SCH28080 ((SCH)E2BeF) in the ribbon representations, as in Fig. 1a. Bound SCH28080 and three Rb⁺ ions are shown as green and purple spheres, respectively. Inset, chemical structure of SCH28080. **b**, Magenta mesh shows anomalous peaks from Rb⁺ contoured at the 5σ level, indicating that three Rb⁺ ions (blue, yellow and red boxes) are bound to the H⁺, K⁺-ATPase (SCH)E2BeF (shown as colour ribbons). Blue, interface between the nucleotide domain and the actuator domain of the symmetry-related neighbouring molecules (grey ribbon). Yellow, K⁺-binding site at the phosphorylation domain, which is homologous to SERCA and Na⁺, K⁺-ATPase. Red, anomalous peak found at the transmembrane cation-binding site. **c**, The molecular surface of (SCH)E2BeF structure, viewed from the luminal side of the membrane. Bound SCH28080 (green sticks) blocks the conduit connecting to the cation-binding site. **d**, Structure as in **c**, but with bound SCH28080 is removed, showing that Rb⁺ bound to the cation-binding site (purple) is exposed to the luminal solution. **e–g**, The Cα traces of the indicated atomic models are superimposed on the H⁺, K⁺-ATPase (Von)E2BeF (blue, with bound vonoprazan shown as spheres).
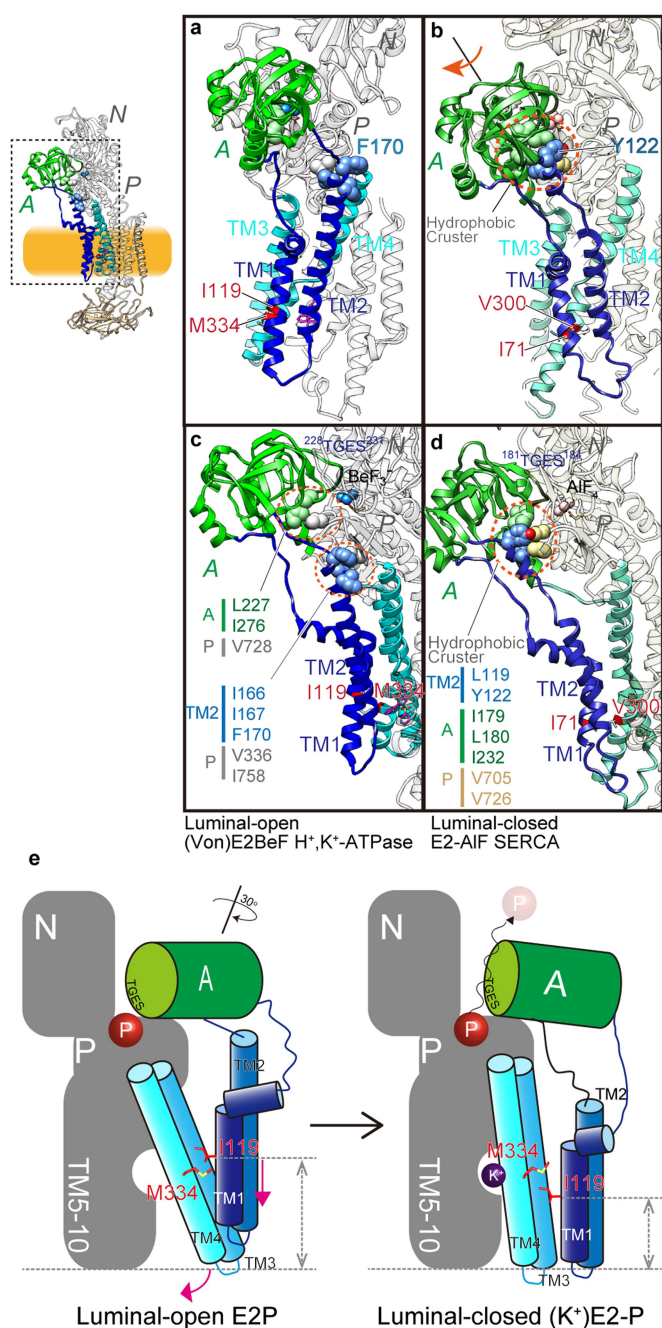
Extended Data Fig. 3 | TM2 helix and the hydrophobic cluster.
a–d, Interface between the actuator and phosphorylation domains, and the cytoplasmic portion of the TM2 in the luminal-open E2P state of H$^+$, K$^+$-ATPase (Von)E2BeF (**a**, **c**) and the luminal-closed E2P transition state of SERCA E2-AlF (PDB code: 2ZBG)[18] (**b**, **d**) are shown. These two atomic models are superimposed according to the TM7–TM10 structure. Broken box on the whole molecular structure (upper left) indicates the region shown in **a–d**, viewed from left (**a**, **b**) or front (**c**, **d**) of the molecule. Actuator domain (green), TM1–TM2 (blue), and TM3–TM4 (cyan) bundles are highlighted. Residues that contribute to the hydrophobic interactions[19] (orange dotted circles) are indicated as spheres with analogous colouring of their respective structural components. Phe170 in H$^+$, K$^+$-ATPase is homologous to Tyr122 in SERCA. Because of the different coordination geometry between phosphate analogues (BeF$_3^-$, light blue; AlF$_4^-$, pink) and the TGES motif (indicated as dark colour in each model) at the interface between the actuator and phosphorylation domains (see Fig. 1e for closed view), the azimuthal position of the actuator domain differs between the two structures (by approximately 30°, as indicated by the orange arrow in **b**). As a consequence, the cytoplasmic portion of TM2 shows different conformations between the α-helical structure in the luminal-open E2P (**a**, **c**) and unwound loop structure in the luminal-closed E2-P forms (**b**, **d**). The C$_\alpha$ positions of Ile119 and Met334 in H$^+$, K$^+$-ATPase (a gating latch) and their homologous residues in SERCA (Ile71 and Val300) are shown in red (see Fig. 3). **e**, Schematic of luminal gate closure in H$^+$, K$^+$-ATPase. In the luminal-open E2P state (left), Ile119 (TM1) and Met334 (TM4) act as a latch to keep the TM1–TM2 bundle in the upright cytoplasmic-side position (indicated by dotted lines and arrows). Binding of counter-transporting K$^+$ to the cation-binding site induces luminal gate closure (right), which is accompanied by the lateral movement of the TM3–TM4 bundle (Fig. 1c) and downward-sliding movement of the TM1–TM2 bundle (indicated by red arrows in the left panel). The sliding movement of TM1–TM2 results in the unwinding of the cytoplasmic portion of TM2 and the rotation of the actuator domain relative to the phosphorylation domain. Finally, bound phosphate at the reaction centre of the phosphorylation domain is hydrolysed owing to the displacement of the TGES loop. Because of the missing interaction between Ile199 and Met334 in their alanine-substituted mutants (Fig. 3c), the TM1–TM2 bundle may slip; therefore, the luminal gate closes spontaneously regardless of K$^+$-binding to the cation-binding site. As a consequence, spontaneous dephosphorylation is induced, producing the K$^+$-independent ATPase activity.
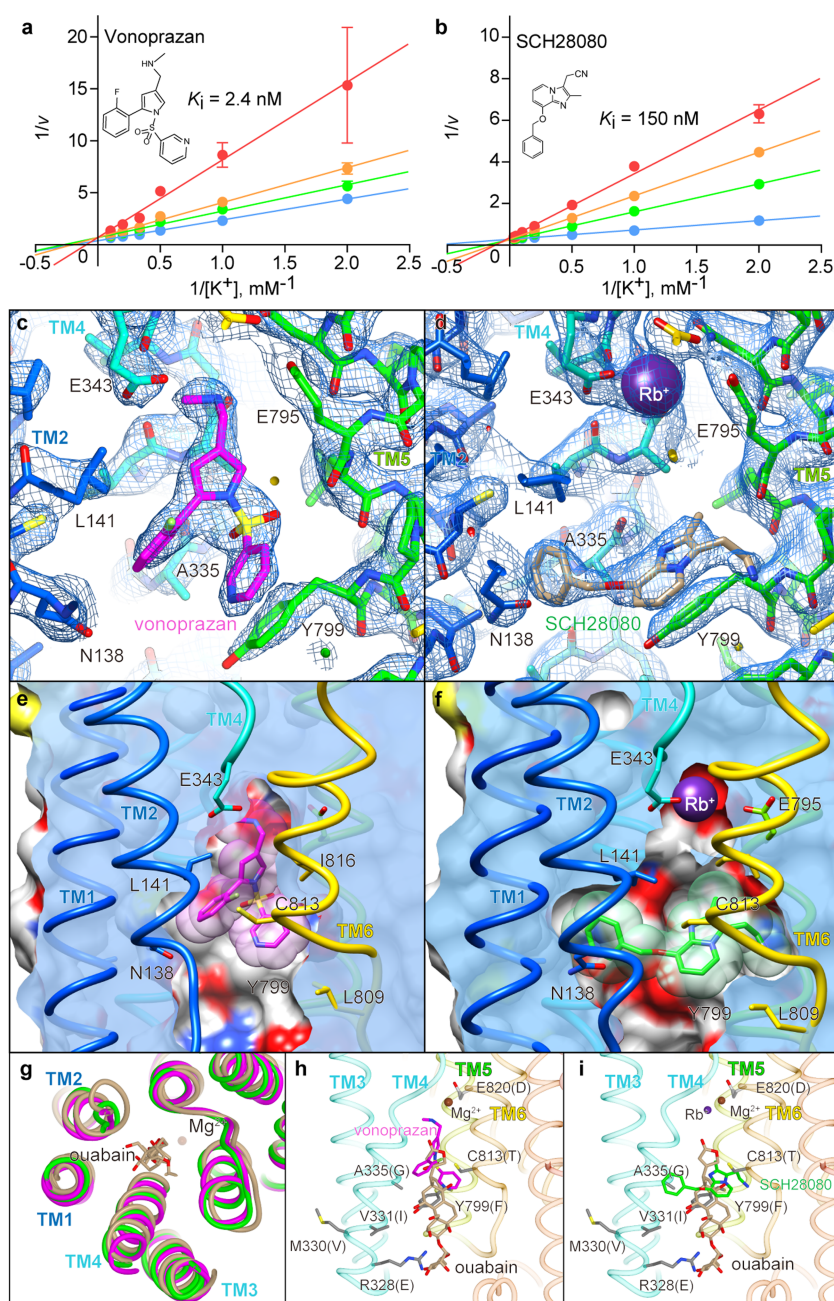
Extended Data Fig. 4 | P-CAB-binding site. Inverse plot of $1/v$ versus $1/[K^+]$ for the wild-type enzyme in the presence of different concentrations of P-CABs (vonoprazan: 0, 5, 10 and 20 nM (**a**); SCH28080: 0, 200, 500 and 1,000 nM (**b**), blue, green, yellow and red circles correspond to the respective P-CAB concentrations), showing typical K^+-competitive inhibition of H^+, K^+-ATPase activity. Data represent mean $\pm$ s.e.m. of triplicated points at each of the indicated K^+ concentrations; representative results from more than three independent measurements are shown. Their chemical structures are provided in each inset. **c, d,** The $2F_o - F_c$ electron density maps (contoured at 2σ) of the vonoprazan- (**c**) and SCH28080-binding site (**d**), viewed from approximately parallel to the membrane plane. In **d**, bound SCH28080 is depicted as wheat colour for clarity. **e, f,** Cross sections of the P-CAB-binding sites perpendicular to the membrane plane. The sectional surface is shown in light blue, and molecular surface is shown as light grey (carbon), with other colours corresponding to different elements (red, oxygen; blue, nitrogen; yellow, sulfur). Transparent spheres for each of the P-CABs represent their van der Waals radius, showing tight binding in their binding pocket. **g,** Structural comparison of the transmembrane region of vonoprazan-bound (magenta), SCH28080-bound (green) H^+, K^+-ATPase and ouabain-bound Na^+, K^+-ATPase (wheat), viewed from luminal side. Bound ouabain and Mg^{2+} ion in the Na^+, K^+-ATPase structure are shown for clarity. **h,** Ouabain and Mg^{2+} ion are superimposed on the vonoprazan-bound H^+, K^+-ATPase structure (ribbons). Seven amino acids of the H^+, K^+-ATPase, for which mutation provides high-affinity ouabain binding, are indicated (grey sticks), and their corresponding amino acids for Na^+, K^+-ATPase are indicated in parentheses. **i,** Bound SCH28080 is superimposed on the structure shown in **h**. See Supplementary Information for details.
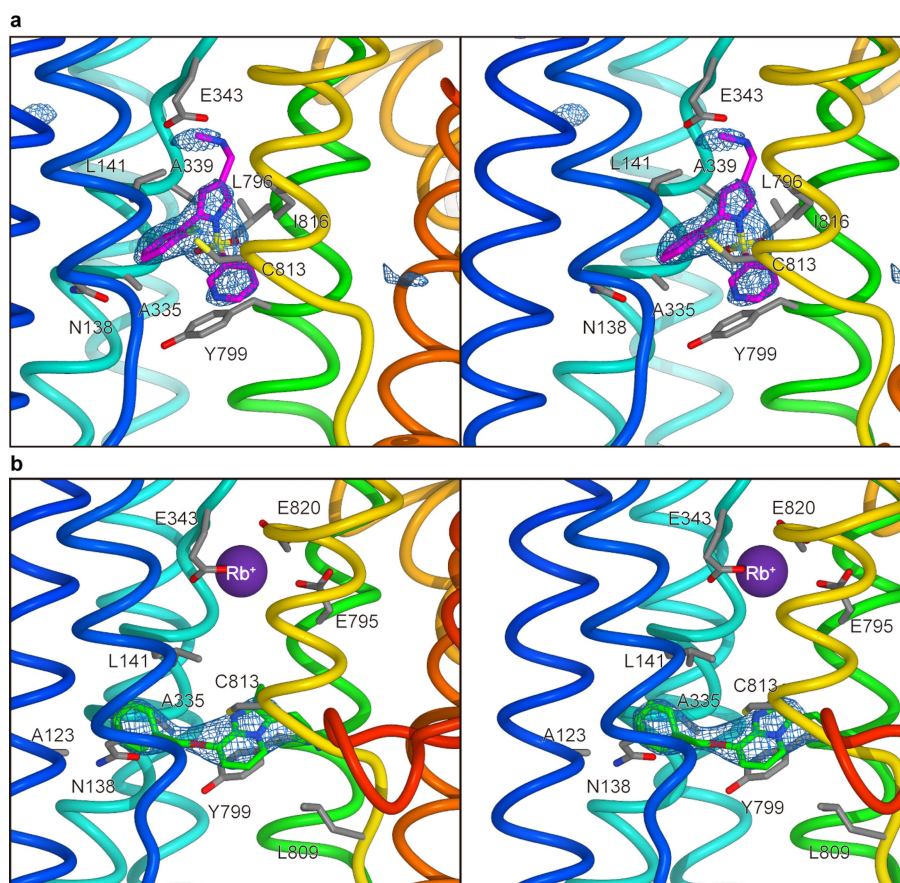
**Extended Data Fig. 5 | $F_o - F_c$ maps for P-CABs.** The $F_o - F_c$ density for vonoprazan (**a**) and SCH28080 (**b**) contoured at 5σ (blue mesh) is shown in stereo view. The amino acids involved in the binding are indicated as sticks.
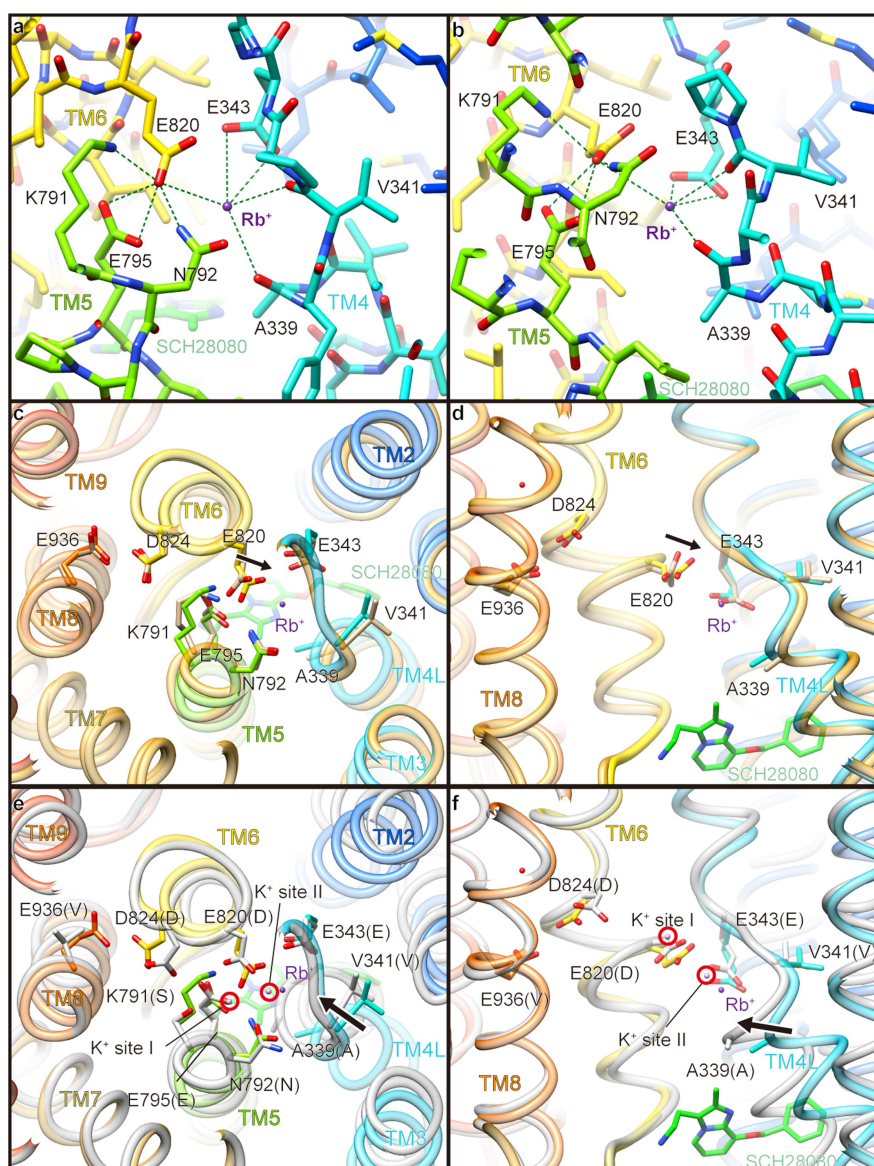
Extended Data Fig. 6 | Cation-binding site in the Rb$^+$-bound, luminal-open E2P state. a, b, Close-up of the cation-binding site in H$^+$, K$^+$-ATPase (SCH)E2BeF viewed approximately perpendicular to the membrane from the cytoplasmic side (**a**) and parallel to the membrane from the TM4 side (**b**). Residues located within 3.5 Å between neighbouring atoms are connected by dotted lines. Bound Rb$^+$ (purple sphere) and water molecules (red) are also indicated. **c, d,** Comparison of the cation-binding site between and Rb$^+$-bound (SCH)E2BeF (colour ribbons) and (Von)E2BeF (wheat), showing the inclination of Glu820 side chain towards Rb$^+$ accompanied by Rb$^+$ binding (arrow). Only polar residues in the observed area are shown for clarity. **e, f,** K$^+$-occluded (K$^+$)$_2$E2-MgF state of Na$^+$, K$^+$-ATPase (light grey, PDB code: 2ZXE) is superimposed on the Rb$^+$-bound (SCH)E2BeF state of H$^+$, K$^+$-ATPase (colour ribbons). Pink spheres highlighted with red circles (site I and II) indicate bound K$^+$ in the Na$^+$, K$^+$-ATPase structure. Atomic models are aligned based on the TM7–TM10 part of the proteins. Arrows indicate displacement of the TM4 luminal portion from the luminal-open to the luminal-closed form. TM5 is removed from the structures shown in **d** and **f** for clarity.
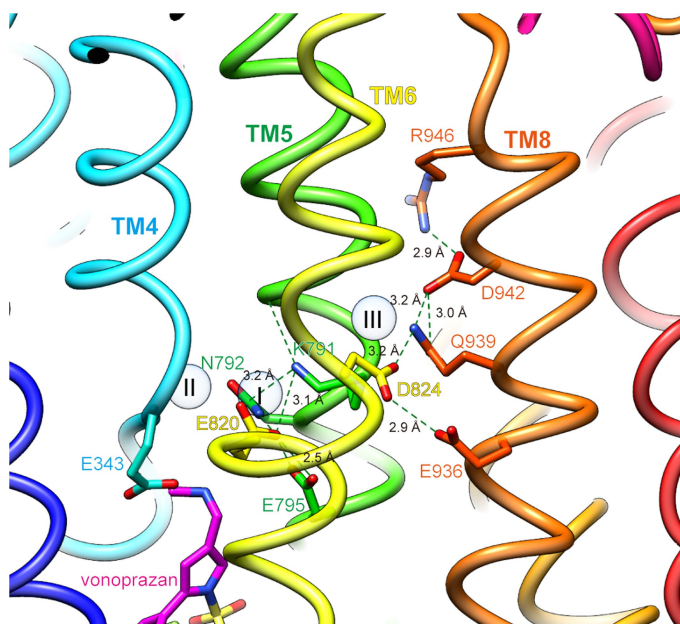
Extended Data Fig. 7 | Hydrogen bond networks. A transmembrane cation-binding site of H^+, K^+-ATPase (Von)E2BeF is shown, viewed from the TM6 side. Only polar residues are shown, and the distances between each residue are provided. Spheres indicate positions responsible for the Na^+-binding site (I–III) in the Na^+, K^+-ATPase E1P-ADP state[43]. The proximity of Asp942 and Arg946 to one another indicates that these residues form a salt bridge.

Extended Data Table 1 | Data collection and refinement statistics

	Vonoprazan (5YLU)	SCH28080 (5YLV)
Data collection		
Space group	$P3_1\,2\,1$	$P3_1\,2\,1$
Cell dimensions		
$\quad a, b, c$ (Å)	104.82, 104.82, 367.08	105.05, 105.05, 368.54
$\quad \alpha, \beta, \gamma$ (°)	90, 90, 120	90, 90, 120
Resolution (Å) [†]	3.2×3.2×2.8 (2.9-2.8)[‡]	3.0×3.0×2.8 (2.9-2.8)
R_{merge}	0.1201(1.689)	0.1186 (2.35)
$I / \sigma I$	11.04 (1.54)	9.88 (0.84)
Completeness (%)	87.51 (38.17)	92.27 (40.65)
Redundancy	5.2 (5.5)	7.8 (8.2)
Refinement		
Resolution (Å)	48.18 – 2.8 (2.9 – 2.8)	48.3 – 2.8 (2.9 – 2.8)
No. reflections	58652 (2197)	59284 (2368)
R_{work} / R_{free}	23.7/28.8 (38.2/42.7)	24.0/29.2 (41.6/47.8)
No. atoms	9884	9938
$\quad$ Protein	9612	9728
$\quad$ Ligand/ion	235	181
$\quad$ Water	37	29
B-factors	43.75	69.82
$\quad$ Protein	43.10	69.39
$\quad$ Ligand/ion	72.45	95.51
$\quad$ Water	28.15	55.10
R.m.s. deviations		
$\quad$ Bond lengths (Å)	0.011	0.010
$\quad$ Bond angles (°)	1.25	1.21

[†]The diffraction data are anisotropic. The resolution limits given are for the $a*$, $b*$ and $c*$ axes, respectively.
[‡]Statistics for the highest-resolution shell are shown in parentheses.

Extended Data Table 2 | ATPase activity of evaluated mutants

	Vonoprazan			SCH28080		
	K_i, nM	x fold	Å	K_i, nM	x fold	Å
Wild-type	2.4 ± 2.3	1	-	150 ± 10	1	-
A123V	3.1 ± 2.3	1.3	7.2	4000 ± 70	27	3.4
N138F	3.5 ± 1.2	1.5	3.7	> 10000	> 67	3.1
L141F	110 ± 41	46	3.5	630 ± 140	4.2	3.6
A335V	12 ± 1.2	5	3.7	> 10000	> 67	3.3
A339S	1.4 ± 0.2	0.6	3.3	300 ± 49	2.0	3.7
E343Q	7.9 ± 1.5	3.3	3.6*	130 ± 21	0.9	8.4
E795Q	17 ± 2.5	7.1	4.3*	600 ± 42	4.0	7.0
L796A	36 ± 5.8	15	3.4	1700 ± 180	11	4.3
Y799A	>500	>200	3.3	> 10000	> 67	3.3
Y799F	9.1 ± 2.7	3.8	3.3	1200 ± 140	8	3.3
L809F	12 ± 6.8	5	4.1	3300 ± 350	22	3.6
C813M	69 ± 15	29	3.3	400 ± 57	2.7	3.6
I816F	150 ± 38	63	3.6	360 ± 52	2.4	4.9

Effect of mutation on the inhibition constant (K_i) for the indicated P-CABs is summarized in the table. K_i value represents the mean ± s.d. determined by K$^+$-competitive inhibition of H$^+$, K$^+$-ATPase activity ($n = 3$, independent experiments). Values indicating fold increase ('x fold') in the K_i value of each mutant, compared with that of wild type, are provided for clarity. Distance between indicated residues and the closest atom of each P-CAB in the crystal structure are also provided. *Distance between the secondary amine of vonoprazan and the closest oxygen atom in the indicated residues is shown. Mutants that considerably affected the affinity of vonoprazan, SCH28080 or both are highlighted in magenta, green or orange, respectively.

Extended Data Table 3 | ATPase activity of mutants

Mutants	V_{max} %	K_m mM	H$^+$-ATPase %
Wild-type	100 ± 3	1.2 ± 0.1	13
I119M	26 ± 1	0.9 ± 0.2	20
M334I	23 ± 1	0.4 ± 0.1	36
I119M/M334I	35 ± 1	0.9 ± 0.1	15
I119A	52 ± 2	-	54
M334A	70 ± 1	-	100
E343D[†]	-	-	-
E343Q	92 ± 3	5.5 ± 0.4	7
E795D	75 ± 1	26 ± 3.2	1
E795Q	125 ± 5	0.4 ± 0.1	30
E820D	90 ± 3	0.75 ± 0.1	9
E820Q	19 ± 2	-	96
D824E	26 ± 2	-	45
D824N	7 ± 0.8	-	73

SCH28080-sensitive ATPase activity of the wild type and indicated mutants (Figs. 3, 4) were determined as in Fig. 3c, and parameters were summarized. Data show the relative amount of the maximum ATPase activity (V_{max}) of mutants compared with the wild-type enzyme (4.3 μmol mg^{-1} h^{-1} in the membrane preparation), K$^+$ affinity (K_m), and fraction of H$^+$-ATPase activity (ATPase activity in the absence of KCl) relative to the maximum H$^+$, K$^+$-ATPase activity of mutants. The specific activity of each mutant enzyme was normalized to the expression level of H$^+$, K$^+$-ATPase determined by fluorescence size-exclusion chromatography. Value represents the mean ± s.e.m. determined by fitting of 24 data points (triplicate of 8 different K$^+$ concentrations) for each experiment.
[†]Data for Glu343Asp mutant are not shown because it exhibited no detectable ATPase activity.

LETTER

Alteration of the magnetosphere of the Vela pulsar during a glitch

Jim Palfreyman[1]*, John M. Dickey[1], Aidan Hotan[2], Simon Ellingsen[1] & Willem van Straten[3]

As pulsars lose energy, primarily in the form of magnetic dipole radiation, their rotation slows down accordingly. For some pulsars, this spin-down is interrupted by occasional abrupt spin-up events known as glitches[1]. A glitch is hypothesized to be a catastrophic release of pinned vorticity[2] that provides an exchange of angular momentum between the superfluid outer core and the crust. This is manifested by a minute alteration in the rotation rate of the neutron star and its co-rotating magnetosphere, which is revealed by an abrupt change in the timing of observed radio pulses. Measurement of the flux density, polarization and single-pulse arrival times of the glitch with high time resolution may reveal the equation of state of the crustal superfluid, its drag-to-lift ratio and the parameters that describe its friction with the crust[3]. This has not hitherto been possible because glitch events happen unpredictably. Here we report single-pulse radio observations of a glitch in the Vela pulsar, which has a rotation frequency of 11.2 hertz. The glitch was detected on 2016 December 12 at 11:36 universal time, during continuous observations of the pulsar over a period of three years. We detected sudden changes in the pulse shape coincident with the glitch event: one pulse was unusually broad, the next pulse was missing (a 'null') and the following two pulses had unexpectedly low linear polarization. This sequence was followed by a 2.6-second interval during which pulses arrived later than usual, indicating that the glitch affects the magnetosphere.

In 2013 we began a three-year observing programme of the Vela pulsar with the aim of recording each single pulse during its next glitch (see Methods). On 2016 December 12 at 11:36 universal time (UT), a glitch of magnitude $\Delta\nu/\nu = 1.431 \times 10^{-6}$ (where $\nu = 11.2$ Hz is the rotation rate) was observed at both the 26-m telescope installed at Mount Pleasant, Tasmania, and the 30-m telescope at Ceduna, South Australia. Extended Data Table 1 shows the arrival times at the Solar System barycentre, as recorded by the two telescopes.

Figure 1 shows a plot of the arrival time residuals of single pulses recorded at Mount Pleasant over a time range of 72 min centred on the glitch. The residuals are the difference between the experimental data and the timing-model results for ν and $\dot{\nu}$, calculated using 36 min of single-pulse data obtained before the glitch.

The inset of Fig. 1 shows a magnification of the plot around the time of the glitch, t_g (vertical red line; see Methods). Near this time, three very-low-probability events occurred: (1) a 'null', which followed an unusually broad pulse, (2) a brief increase in the mean of the timing residuals, implying either a decrease in ν or, more probably, a change in the magnetosphere that affected timings, and (3) a reduction in the variance of the timing residuals.

Figure 2 shows 11 consecutive pulses including the 'null' that occurred at pulse number 77 (in the recorded file). Although pulses 72–75 look typical, pulse 76 looks different: the flux is spread smoothly over about 10 ms, the entire width of the integrated pulse profile of the Vela pulsar. We have not seen a similarly broad pulse shape in the more than 100,000 pulses that we have examined.

The pulse following this broad pulse is the 'null' pulse, and pulses 78 and 79 show minimal linear polarization, as demonstrated by the

absence of a position angle swing (right column of Fig. 2). Then, typical pulse shapes are again observed from pulse 80 onwards. Analysis of data collected on other days shows that on average, the single-pulse flux density is below the detection threshold of the 26-m telescope once every 77,700 pulses.

Although some pulsars show frequent null pulses, Vela does not[4,5], and general pulsar observations indicate that nulls are not expected to occur in young pulsars such as Vela[6]. We cannot determine whether pulse 77 in Fig. 2 is a true null, with zero flux emitted, a very faint pulse
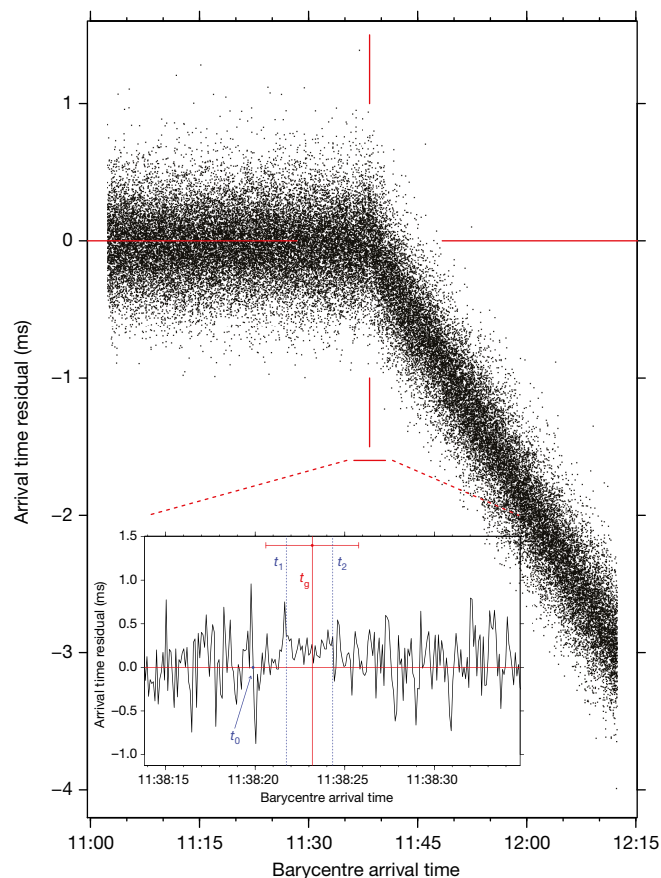

Fig. 1 | Timing residuals of single pulses near the time of the glitch. The horizontal axis shows the arrival time at the Solar System barycentre on modified Julian day 57,734, and the vertical axis shows the residual of the arrival time, obtained from the pre-glitch model. The vertical red line marks the fitted time of the glitch (t_g). The inset shows a magnification of the plot. 3.3 s before t_g, a 'null' occurred (t_0), followed by an unusual change in the timing residuals, with late mean arrival times and reduced variances. Because the 'null' cannot be timed, it has been placed on the 0.0 ms line. The horizontal error bar represents the 1σ uncertainty in the fitting of t_g.

[1]University of Tasmania, Sandy Bay, Tasmania, Australia. [2]CSIRO Astronomy and Space Science, Kensington, Western Australia, Australia. [3]Auckland University of Technology, Auckland, New Zealand. *e-mail: jim77742@gmail.com
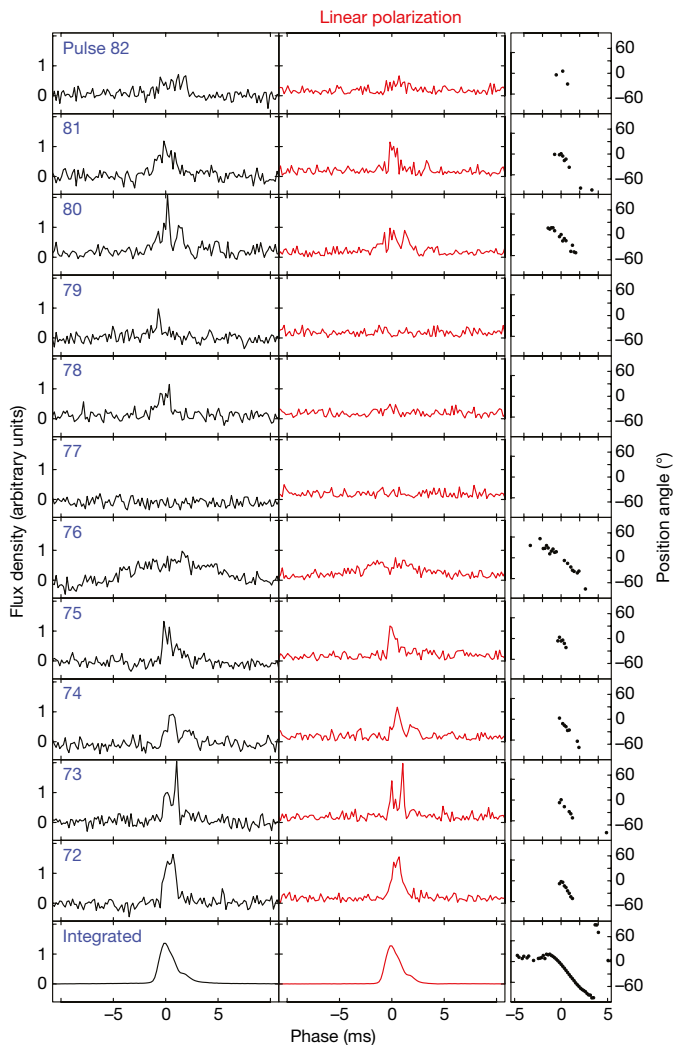
Linear polarization

Fig. 2 | A contiguous sequence of single pulses surrounding the 'null'. Each row corresponds to a single pulse, with time increasing from bottom to top and the pulse number (in the recorded file) indicated in blue. The 'null' is pulse 77. For reference, the bottom row shows the integrated pulse profile. The left panels show the total flux density in arbitrary units, the middle panels show linear polarization and the right panels show the position angle of the linear polarization. Circular polarization was negligible and is not shown. The slight offset in the linear polarization is due to off-pulse noise. Only about a fifth of the pulse period is shown. The position angle is not plotted for pulses 78 and 79 because no linear polarization was detected immediately after the 'null'.

that is below the detection threshold of the 26-m telescope, or even a pulse with more severe broadening than pulse 76. However, such a pulse is a rare event. The 'null' pulse appears at time t_0, only 3.3 s (37 pulsar rotations) before the best estimate of t_g, which has a 1σ uncertainty of 2.5 s. The probability of a null appearing anywhere in the 37 rotations before the glitch is $P = 4.8 \times 10^{-4}$.

Soon after the 'null', at $t_1 = t_0 + 1.8$ s (20 pulsar rotations), a substantial change occurred in both the mean and the variance of the timing residuals, which lasted for 2.6 s (29 pulsar rotations), until time t_2. We searched two other full days of data (more than about 1.4×10^6 pulses) for a sequence of pulses of similar length and with a greater change in the mean, combined with a smaller change in variance than that observed here. None was found. Figure 3 shows a scatter plot of the mean and standard deviation (σ) of single pulses over the 36-min period before t_g, as shown in the left half of Fig. 1. This extraordinary offset in the mean arrival times of the sequence of pulses and the low corresponding variance suggest that the pulsar emission mechanism was affected by the glitch process during this interval.
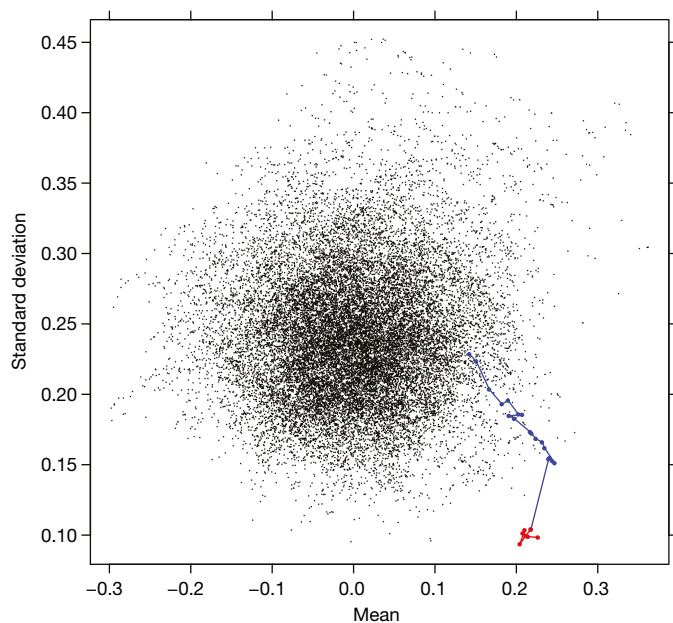

Fig. 3 | Scatter plot of the mean and standard deviation of single-pulse timing residuals. Data are shown for the 36 min leading up to the glitch (left half of Fig. 1), calculated using a sliding window of 21 data points. The blue dots correspond to the period t_0–t_1 and the red outliers to t_1–t_2. The connecting lines show how the sequence progresses. The units are milliseconds.

Figure 4a shows a 260-s view of the timing residuals, with the 'null' at t_0 marked, Fig. 4b provides the cumulative sum of the timing residuals, and Fig. 4c shows the cumulative sum after glitch modelling has been applied to the 72 min of data. The cumulative sums highlight overall changes that are not apparent in the residual plot. The sequence of pulses showing increased mean and reduced variance commences at t_1 and finishes at t_2. Label t_3 marks what appears to be a permanent speed-up in rotation after the glitch process has been completed.

We note that t_g can be fitted to a precision of only 2.5 s, but the 'null' pulse provides a fiducial time t_0 with a precision of the pulsar rotation rate, 89 ms. The timing of the spin-down, from t_1 to t_2, is based on the sustained change in the mean and variance shown in the inset of Fig. 1. Extended Data Table 2 shows the arrival times of these events at the Solar System barycentre.

The 2.6 s from t_1 to t_2 could be associated with the unpinning process of superfluid vortices, and the associated changes in angular momentum, which are presumed to be the cause of pulsar glitches. An alternative explanation is changes in the magnetosphere triggered by the glitch. These changes could be caused by the unpinning of the vortices affecting the magnetic flux tubes in the core.

The 4.4-s interval (49 pulsar rotations) between t_0 and t_2 may indicate the rise time (τ_r) of the glitch, that is, the time required to transfer angular momentum from the superfluid-permeated inner crust to the outer crust. The rise time of the glitch has implications for the equation of state. Sourie et al.[3] compare the predictions of two equations of state, the density-dependent hadronic (DDH) model and DDHδ, which takes into account a scalar isovector interaction channel. For a pulsar mass of $1.3\,M_\odot$–$1.6\,M_\odot$, where $M_\odot$ is the mass of the Sun, the DDH model predicts a glitch rise time of 4–5.5 s and DDHδ predicts 2.5–3.5 s. If τ_r is indeed 4.4 s, then DDH might be the preferred equation-of-state model.

The 43.8-s interval (490 pulsar rotations) between t_2 and t_3 may correspond to the time after the glitch when the crust and interior are synchronized, before their rotation rates become decoupled.

Sedrakian & Cordes[7] present a model in which the crustal magnetic field provides a potential barrier against the superconducting proton vortices in the core, which in turn act as a barrier to the superfluid vortices that are trying to migrate outwards. On the basis of this model,
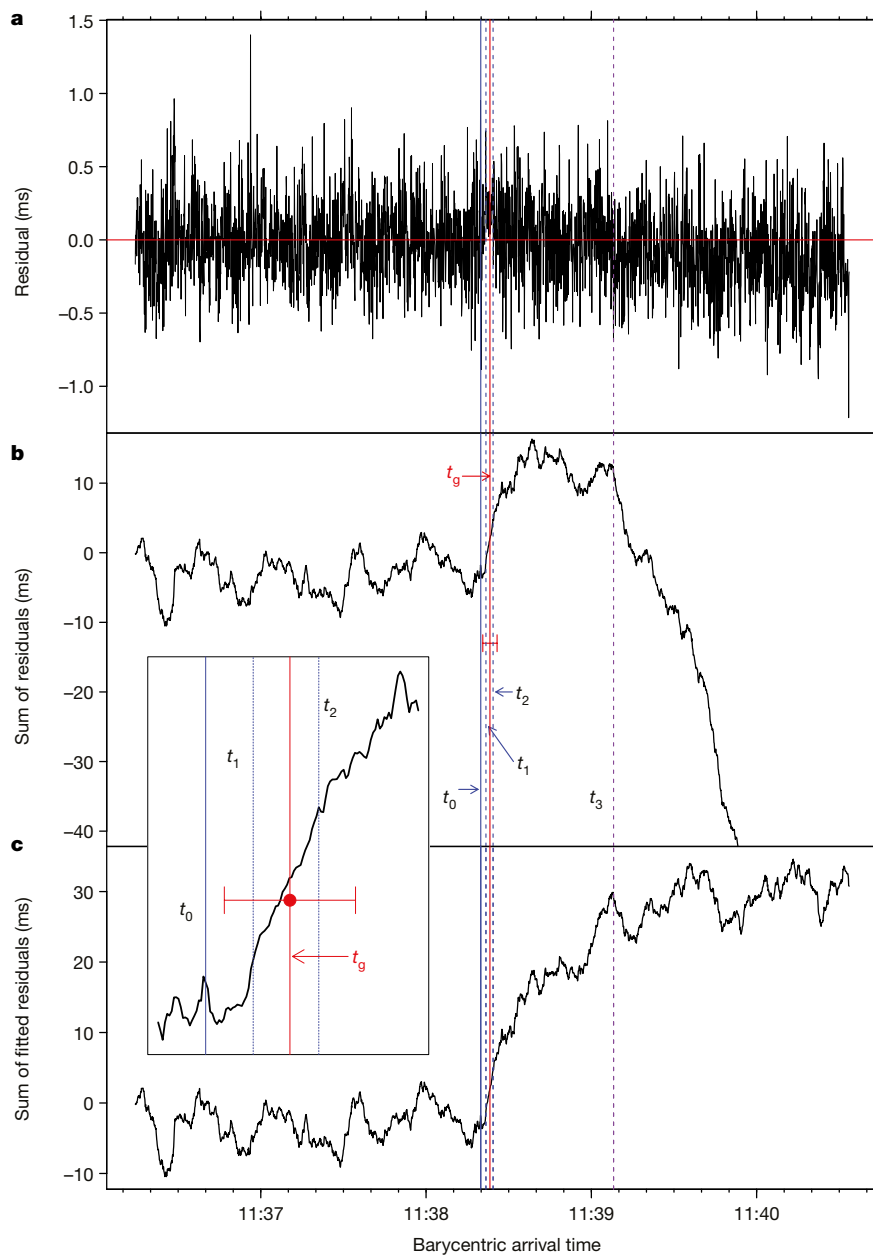
Fig. 4 | Timing residuals and their cumulative sum around the time of the glitch. Residuals are shown for the 260 s around the time of the glitch t_g (solid red line). **a**, Timing residuals (in milliseconds) similar to those of Fig. 1, with no glitch modelling applied. **b**, Cumulative sum of the timing residuals of **a**. **c**, Cumulative sum of timing residuals, after glitch modelling has been applied. The events observed at times t_0–t_3 (see text) are highlighted. Inset, magnified view of **b** showing t_0, t_1, t_g and t_2. The horizontal error bar represents the 1σ uncertainty in the fitting of t_g.
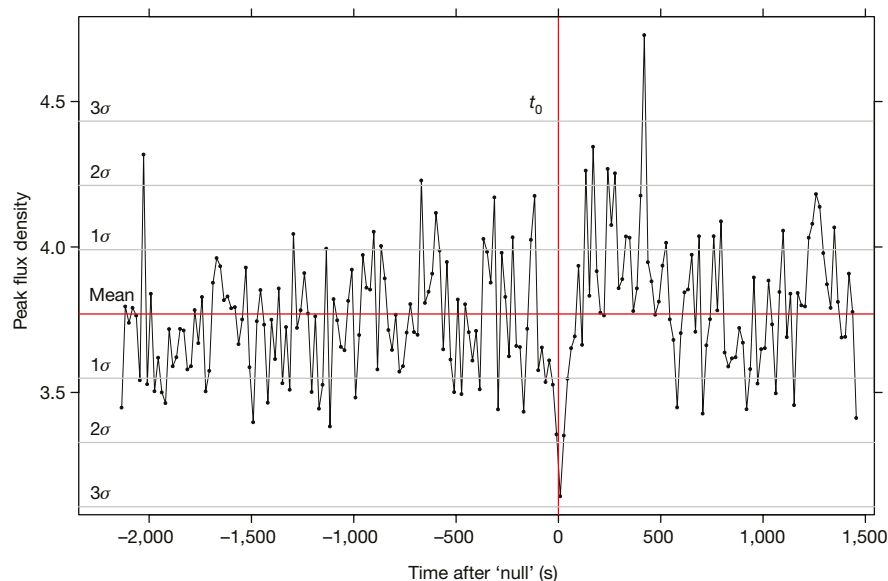

Fig. 5 | Peak flux density around the time of the 'null'. The flux density is shown in arbitrary units and the 'null' occurs at t_0 (vertical red line). Data have been binned into 200-pulse (about 18 s) bins. The horizontal lines indicate 1σ spacings.

they predict that a glitch would affect the geometry of the pulsar's magnetic field. This may be what we have observed in the 'null' pulse (pulse 77), the strange shape of pulse 76, and the loss of linear polarization in pulses 78 and 79.

We also observed a 3σ dip in the peak flux density for about 2 min on either side of t_0 (see Fig. 5). Vela is known[8] to emit bright pulses that arrive between 1 ms and 1.5 ms before the main pulse. This 3σ dip, combined with the reduced variance of the timing residuals, suggests that fewer bright pulses were emitted from the magnetosphere in this interval. The disruption of the magnetosphere could have caused the normal coherent emission process to break down sufficiently to stop the emission of bright pulses from the precursor region, where they are usually seen. Changes in the particle bunching in the magnetosphere could affect coherence, the radio flux density, the beaming direction or the emission height.

Future observations of single pulses associated with glitches in Vela may provide confirmation that glitches consistently cause null pulses or peculiar-shaped pulses. Observations with larger telescopes (or telescope arrays) may probe this behaviour more deeply by determining whether the 'null' is genuine, which will help us to resolve some of the outstanding issues with regard to the internal mechanics and equations of state of neutron stars.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0001-x.

1. Rees, M. J. & Trimble, V. L. Physical sciences: planet, pulsar, "glitch" and wisp. *Nature* **229**, 395–396 (1971).
2. Anderson, P. W. & Itoh, N. Pulsar glitches and restlessness as a hard superfluidity phenomenon. *Nature* **256**, 25–27 (1975).
3. Sourie, A., Chamel, N., Novak, J. & Oertel, M. Global numerical simulations of the rise of vortex-mediated pulsar glitches in full general relativity. *Mon. Not. R. Astron. Soc.* **464**, 4641–4657 (2017).
4. Biggs, J. D. An analysis of radio pulsar nulling statistics. *Astrophys. J.* **394**, 574–580 (1992).
5. Johnston, S., van Straten, W., Kramer, M. & Bailes, M. High time resolution observations of the Vela pulsar. *Astrophys. J.* **549**, L101–L104 (2001).
6. Rankin, J. M. Toward an empirical theory of pulsar emission. III - mode changing, drifting subpulses, and pulse nulling. *Astrophys. J.* **301**, 901–922 (1986).
7. Sedrakian, A. & Cordes, J. M. Vortex-interface interactions and generation of glitches in pulsars. *Mon. Not. R. Astron. Soc.* **307**, 365–375 (1999).
8. Krishnamohan, S. & Downs, G. S. Intensity dependence of the pulse profile and polarization of the Vela pulsar. *Astrophys. J.* **265**, 372–388 (1983).

Reviewer information *Nature* thanks J. Cordes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.P. conducted all the observations, led the analysis and the writing of the manuscript and produced all the plots. J.M.D. provided input on the manuscript and analysis. W.v.S. conducted the polarization calibration and assisted with the manuscript and analysis. S.E. and A.H. assisted with the manuscript and analysis.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0001-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Using the Mount Pleasant 26-m radio telescope, which is located near Hobart, Tasmania, we observed Vela when it was above the lower elevation limit, 4.3°, obtaining data for about 19 h each day. We also observed Vela with our 30-m telescope in Ceduna, South Australia. Both telescopes operated at a centre frequency of 1,376 MHz and a bandwidth of 64 MHz. Although the Ceduna dish is larger than that of the Mount Pleasant 26-m telescope, its receiver is much less sensitive because it is not cooled to cryogenic temperatures. Both telescopes have dual linearly polarized receivers.

We recorded about 14,000 h of baseband voltage data from Mount Pleasant in both polarizations at a rate of 128×10^6 samples per second. Data from the Ceduna telescope were recorded in a buffer and discarded until the glitch occurred.

The baseband data files from both observatories were coherently de-dispersed, detected and integrated into single pulses using DSPSR[9]. In the time domain, each rotation of the pulsar was divided into 8,192 phase intervals (giving a resolution of 10.9 μs) and in the frequency domain, the 64-MHz band was divided into 16 sub-bands. PSRCHIVE[10] was used for analysis, and polarization calibration was performed by using Vela as a polarized reference source, but using 128 frequency sub-bands and 1,024 phase intervals[11].

The glitch epoch (t_g) was calculated using the TEMPO2 software[12,13] and a two-stage iterative process. First, we adjusted t_g to minimize the phase ($\Delta\phi$). We modelled for changes in ν and $\dot\nu$ and set the long-term glitch decay parameters[14] to $\Delta\nu_d = 1.29 \times 10^{-7}$ and $\tau_d = 0.96$. Then, we used an iterative process and stopped when $\Delta\phi < 10^{-7}$ ($\Delta\phi = 6.98 \times 10^{-8} \approx 6$ ms).

After this approximation, we adjusted t_g manually to minimize the root-mean-square residuals in the arrival time (data minus model). Then, we adjusted ν to minimize the root-mean-square residuals, and then $\dot\nu$. This was repeated several times, until convergence was achieved. In each step of this process, the plot of the root-mean-square residuals was a parabola smooth enough to validate our best-fit determination.

The $\Delta\nu/\nu$ and $\Delta\dot\nu/\dot\nu$ values for the Mount Pleasant observations shown in Extended Data Table 1 were obtained using four days of data, whereas the fitting of the glitch epoch was based on 72 min of data. The corresponding results for Ceduna were based on 15 h of data, but with only about 1 h of pre-glitch timings available; thus, $\Delta\nu/\nu$ was not well constrained and $\Delta\dot\nu/\dot\nu$ could not be determined.

Data availability. Source Data files containing the data shown in the figures are available in the online version of the paper. The raw data were generated at the Mount Pleasant and Ceduna radio observatories, which are operated by the University of Tasmania, and are available from the corresponding author upon reasonable request.

Code availability. The software DSPSR, TEMPO2 and PSRCHIVE are available at http://dspsr.sourceforge.net/, http://www.atnf.csiro.au/research/pulsar/tempo2/ and http://psrchive.sourceforge.net/, respectively.

9. van Straten, W. & Bailes, M. DSPSR: digital signal processing software for pulsar astronomy. *Publ. Astron. Soc. Aust.* **28**, 1–14 (2011).
10. Hotan, A. W., van Straten, W. & Manchester, R. N. PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis. *Publ. Astron. Soc. Aust.* **21**, 302–309 (2004).
11. van Straten, W. High-fidelity radio astronomical polarimetry using a millisecond pulsar as a polarized reference source. *Astrophys. J. Suppl. Ser.* **204**, 13 (2013).
12. Hobbs, G. B., Edwards, R. T. & Manchester, R. N. TEMPO2, a new pulsar-timing package–I. An overview. *Mon. Not. R. Astron. Soc.* **369**, 655–672 (2006).
13. Edwards, R. T., Hobbs, G. B. & Manchester, R. N. TEMPO2, a new pulsar timing package–II. The timing model and precision estimates. *Mon. Not. R. Astron. Soc.* **372**, 1549–1574 (2006).
14. Sarkissian, J. M., Reynolds, J. E., Hobbs, G. & Harvey-Smith, L. One year of monitoring the Vela pulsar using a phased array feed. *Publ. Astron. Soc. Aust.* **34**, e027 (2017).

Extended Data Table 1 | Arrival times of the 2016 glitch of the Vela pulsar

Location	MJD	Time (UTC)	$\frac{\Delta\nu}{\nu}$	$\frac{\Delta\dot{\nu}}{\dot{\nu}}$
Mount Pleasant	57734.484991	11:38:23.2	1431.24×10^{-9}	9.2×10^{-3}
uncertainty	$\pm 2.9 \times 10^{-5}$	± 2.5 s	$\pm 0.069 \times 10^{-9}$	$\pm 0.83 \times 10^{-3}$
Ceduna	57734.484973	11:38:21.7	1433.5×10^{-9}	N/A
uncertainty	$\pm 3.2 \times 10^{-5}$	± 2.8 s	N/A	N/A

Arrival times at the Solar System barycentre were estimated on the basis of data recorded at each observatory. The last two columns list the relative change in rotation frequency and the relative change in the first derivative of the rotation frequency. Uncertainties are 1σ. MJD, modified Julian date; UTC, coordinated universal time.

Extended Data Table 2 | Arrival times of key events at the Solar System barycentre

	Event	MJD	Time	Δt (s)	Rotations
t_0	null pulse	57734.4849521	11:38:19.9	1.8	20
t_1	spin-down starts	57734.4849738	11:38:21.7	1.5	17
t_g	glitch fit	57734.4849906	11:38:23.2	1.1	12
t_2	spin-down ends	57734.4850038	11:38:24.3	43.8	490
t_3	spin-up starts	57734.48551	11:39:08.1		

The times t_g and t_0–t_3 are listed, as shown in Fig. 4. The last two columns list the time difference (Δt) and number of pulsar rotations between events. MJD, modified Julian date.

LETTER

Experimentally generated randomness certified by the impossibility of superluminal signals

Peter Bierhorst[1,2]*, Emanuel Knill[1,3], Scott Glancy[1], Yanbao Zhang[1,8], Alan Mink[4,5], Stephen Jordan[4], Andrea Rommal[6], Yi-Kai Liu[4], Bradley Christensen[7], Sae Woo Nam[1], Martin J. Stevens[1] & Lynden K. Shalm[1,2]

From dice to modern electronic circuits, there have been many attempts to build better devices to generate random numbers. Randomness is fundamental to security and cryptographic systems and to safeguarding privacy. A key challenge with random-number generators is that it is hard to ensure that their outputs are unpredictable[1–3]. For a random-number generator based on a physical process, such as a noisy classical system or an elementary quantum measurement, a detailed model that describes the underlying physics is necessary to assert unpredictability. Imperfections in the model compromise the integrity of the device. However, it is possible to exploit the phenomenon of quantum non-locality with a loophole-free Bell test to build a random-number generator that can produce output that is unpredictable to any adversary that is limited only by general physical principles, such as special relativity[1–11]. With recent technological developments, it is now possible to carry out such a loophole-free Bell test[12–14,22]. Here we present certified randomness obtained from a photonic Bell experiment and extract 1,024 random bits that are uniformly distributed to within 10^{-12}. These random bits could not have been predicted according to any physical theory that prohibits faster-than-light (superluminal) signalling and that allows independent measurement choices. To certify and quantify the randomness, we describe a protocol that is optimized for devices that are characterized by a low per-trial violation of Bell inequalities. Future random-number generators based on loophole-free Bell tests may have a role in increasing the security and trust of our cryptographic systems and infrastructure.

The search for certifiably unpredictable random-number generators is motivated by applications, such as secure communication, for which the predictability of pseudorandom strings makes them unsuitable. Private randomness is required to initiate and authenticate virtually every secure communication[15], and public randomness from randomness beacons can be used for public certification and resource distribution in many settings[16]. To certify randomness, we can perform an experiment known as a Bell test[17]; in its simplest form, the Bell test involves performing measurements on an entangled system with components located in two physically separated measurement stations, where at each station a choice is made between one of two types of measurement. After multiple experimental trials with varying measurement choices, if the measurement data violate conditions known as 'Bell inequalities', then the data are certified to contain randomness under weak assumptions.

Our randomness generation uses a 'loophole-free' Bell test, which is characterized by high detection efficiency and space-like separation of the measurement stations during each experimental trial. The bits are unpredictable assuming (1) that the choices of measurement setting are independent of the experimental devices and of pre-existing classical information about them and (2) that, in each experimental trial, the measurement outcomes at each station are independent of the settings

at the other station. The first assumption is ultimately untestable, but the premise that it is possible to choose measurement settings independently of a system being measured is often tacitly invoked in the interpretation of many scientific experiments and laws of physics[18]. The second assumption can be violated only if signals can be sent faster than the speed of light, given our trust that the space-like separation of the relevant events in the experiment is accurately verified by the timing electronics and that the results are final when recorded. We also trust that the classical computing equipment used to process the data operates according to specification.

Under the above assumptions, the output randomness is certified to be unpredictable with respect to a real or hypothetical actor 'Eve', who is in possession of the pre-existing classical information, is physically isolated from the devices while they are under our control and is without access to data produced during the protocol. The bits remain unpredictable to Eve if she learns the settings at any time after her last interaction with the devices. If the devices are trusted, which is reasonable if we built them, then this final interaction may be well before the start of the protocol, in which case the settings can come from public randomness[2,10]. In particular, an existing public randomness source can be used, such as the National Institute of Standards and Technology (NIST) random beacon[16], to generate much-needed private randomness as output. Because the assumptions do not constrain the specific physical realization of the devices and do not require specific states or measurements, they implement a 'device-independent' framework[4,19,20], which allows an individual user to assure security with minimal assumptions about the devices.

Compared to other implementations of random-number generations that invoke device-independence[5,21], our implementation is notable because it enforces space-like separation between measurement stations. Bell tests that achieve space-like separation without other experimental loopholes have been performed only recently[12–14,22]. It can be argued that interaction between spatially (if not space-like) separated measuring stations can be assumed to be negligible. However, any shielding between the stations is necessarily incomplete; for example, there must be an open quantum channel to establish entanglement. Mundane physical effects, such as accidentally scattered photons, can allow predictable systems to appear to violate Bell inequalities when shielding is incomplete. Relying instead on the impossibility of faster-than-light communication provides stronger assurance of the unpredictability of the randomness.

We generated randomness using an improved version of a recently reported[13] loophole-free Bell test (which was subsequently used elsewhere[23]). We collected five datasets, with the best-performing one yielding 1,024 random bits that are uniformly distributed to within 10^{-12}, as measured by the total variation distance (see below). We also obtained 256 random bits from the main dataset analysed previously[13], albeit uniform only to within 0.02; see Supplementary Information section 6. The experiment, illustrated in

[1]National Institute of Standards and Technology, Boulder, CO, USA. [2]Department of Physics, University of Colorado, Boulder, CO, USA. [3]Center for Theory of Quantum Matter, University of Colorado, Boulder, CO, USA. [4]National Institute of Standards and Technology, Gaithersburg, MD, USA. [5]Theiss Research, La Jolla, CA, USA. [6]Muhlenberg College, Allentown, PA, USA. [7]Department of Physics, University of Wisconsin, Madison, WI, USA. [8]Present address: NTT Basic Research Laboratories and NTT Research Center for Theoretical Quantum Physics, NTT Corporation, Atsugi, Japan. *e-mail: peter.bierhorst@nist.gov
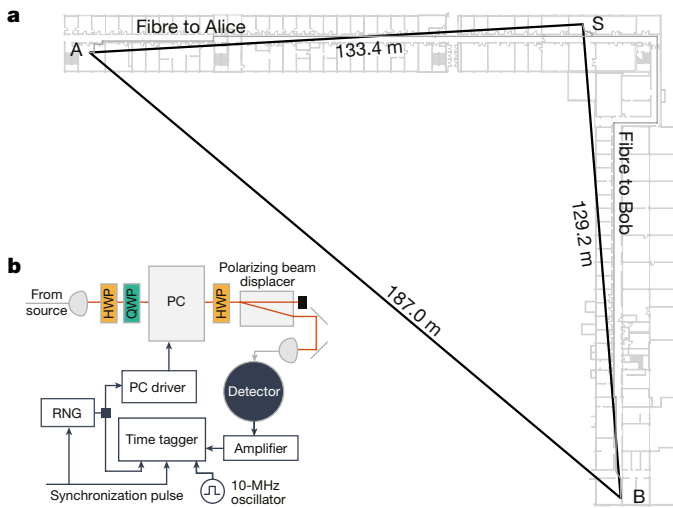
Fig. 1 | Diagram of the experiment. a, b, The relative locations of the source (S), Alice (A) and Bob (B) are depicted in **a**. In each trial, the source laboratory produces a pair of photons in a non-maximally polarization-entangled state. One photon is sent to Alice's laboratory while the other is sent to Bob's laboratory to be measured, as shown in **b**. Alice and Bob both use a fast Pockels cell (PC), two half-wave plates (HWPs), a quarter-wave plate (QWP) and a polarizing beam displacer to switch between their respective polarization measurements. A pseudorandom-number generator (RNG) governs the choice of each measurement setting for each trial. After passing through the polarization optics, the photons are sent to a superconducting nanowire detector. The signals from the detector are amplified and sent to a time tagger, where their arrival times are recorded and the measurement outcome is fixed. Alice's measurement outcome is space-like separated from the triggering of Bob's Pockels cell and vice versa.

Fig. 1, consisted of a source of entangled photons and two measurement stations, named 'Alice' and 'Bob'. During an experimental trial, at each station a random choice was made between two measurement settings, labelled 0 and 1, after which a measurement outcome of detection (+) or non-detection (0) was recorded. Each station's implementation of the measurement setting was space-like separated from the other station's measurement event, and no post-selection was used in collecting the data; see Methods for details. For trial i, we model Alice's settings choices with the random variable X_i and Bob's with Y_i, both of which take values in the set $\{0, 1\}$. Alice's and Bob's measurement-outcome random variables are A_i and B_i, respectively, both of which take values in the set $\{+, 0\}$. When referring to a generic single trial, we omit the i indices. With this notation, a general Bell inequality for our scenario can be expressed in the form[24]

$$\sum_{abxy} s_{xy}^{ab} \mathbb{P}(A = a, B = b \mid X = x, Y = y) \le \beta \qquad (1)$$

where s_{xy}^{ab} are fixed real coefficients indexed by a, b, x and y, which range over all possible values of A, B, X and Y, and $\mathbb{P}$ denotes probability. The upper bound β is required to be satisfied whenever the settings-conditional outcome probabilities are induced by a model that satisfies 'local realism'. Local-realist distributions, which cannot be certified to contain randomness, are those for which $\mathbb{P}(A = a, B = b \mid X = x, Y = y)$ is of the form $\sum_\lambda \mathbb{P}(A = a \mid X = x, \Lambda = \lambda)\mathbb{P}(B = b \mid Y = y, \Lambda = \lambda)\mathbb{P}(\Lambda = \lambda)$ for a random variable Λ that represents local hidden variables. The Bell inequality is non-trivial if there exists a quantum-realizable distribution that can violate the bound β.

It has long been known that experimental violations of Bell inequalities such as equation (1) indicate the presence of randomness in data. To quantify randomness with respect to Eve, we represent Eve's initial classical information by a random variable E. We formalize the assumption

that measurement settings can be generated independently of the system being measured and of Eve's information with the following condition:

$$\mathbb{P}(X_i = x, Y_i = y \mid E = e, \text{past}_i) = \mathbb{P}(X_i = x, Y_i = y)$$
$$= \frac{1}{4} \quad \forall x, y, e \qquad (2)$$

where past_i represents events prior to the ith trial, specifically including the trial settings and outcomes for trials 1 to $i - 1$. Our other assumption, that measurement outcomes are independent of remote measurement choices, is formalized as follows:

$$\mathbb{P}(A_i = a \mid X_i = x, Y_i = y, E = e, \text{past}_i)$$
$$= \mathbb{P}(A_i = a \mid X_i = x, E = e, \text{past}_i) \quad \text{and}$$
$$\mathbb{P}(B_i = b \mid X_i = x, Y_i = y, E = e, \text{past}_i)$$
$$= \mathbb{P}(B_i = b \mid Y_i = y, E = e, \text{past}_i) \quad \forall a, b, x, y, e \qquad (3)$$

These equations are commonly referred to as the 'non-signalling' assumptions, although they are often stated without the conditionals E and past_i. Our space-like separation of settings and remote measurements provide assurance that the experiment obeys equation (3). If we were to assume that the measured systems obey quantum physics, then stronger constraints are possible[25,26].

Given equations (2) and (3), our protocol produces random bits in two sequential parts. For the first part, 'entropy production', we implement n trials of the Bell test, from which we compute a statistic V that is related to a Bell inequality (equation (1)). V quantifies the Bell violation and determines whether or not the protocol passes or aborts. If the protocol passes, then we certify an amount of randomness in the outcome string whether or not Eve has access to the setting string. In the second part, 'extraction', we process the outcome string into a shorter string of bits, the distribution of which is close to uniform. We used our customized implementation of the Trevisan extractor[27] derived from the framework of Mauerer, Portmann and Scholz[28] and the associated open-source code. We call this the TMPS algorithm; see Supplementary Information section 4 for details.

We applied a new method of certifying the amount of randomness in Bell tests. Previous methods for related models with various sets of assumptions[2-8,29,30] are ineffective in our experimental regime (see Supplementary Information section 7), which is characterized by a small per-trial violation of Bell inequalities. Other recent works that explore ways of effectively certifying randomness from a wider range of experimental regimes assume that measured states are independent and identically distributed (i.i.d.) or that the regime is asymptotic[9-11,31]. Our method, which does not require these assumptions, builds on the prediction-based ratio method for rejecting local realism[32]. Applying this method to training data (see below), we obtain a real-valued Bell function T with arguments A, B, X and Y that satisfies $T(A, B, X, Y) > 0$ with expectation $\mathbb{E}(T) \le 1$ for any local-realist distribution that satisfies equation (2). From T we determine the maximum value $1 + m$ of $\mathbb{E}(T)$ over all distributions that satisfy equations (2) and (3), where we require that $m > 0$. Such a function T induces a Bell inequality (equation (1)) with $\beta = 4$ and $s_{xy}^{ab} = T(a, b, x, y)$. Define $T_i = T(A_i, B_i, X_i, Y_i)$ and $V = \prod_{i=1}^n T_i$; if the experimenter observes a value of V larger than 1, this indicates a violation of the Bell inequality and the presence of randomness in the data. The randomness is quantified by the 'entropy production theorem' (see below), which we prove in Supplementary Information section 2. We denote all of the settings of both stations with $\mathbf{XY} = X_1 Y_1 X_2 Y_2 \ldots X_n Y_n$; other sequences such as $\mathbf{AB}$ and $\mathbf{ABXY}$ are similarly interleaved over n trials.

The entropy production theorem is as follows. Suppose T is a Bell function that satisfies the above conditions. Then, in an experiment of n trials that obey equations (2) and (3), the following inequality holds for all $\epsilon_p \in (0, 1)$ and v_{thresh} satisfying $1 \le v_{\text{thresh}} \le [1 + (3/2)m]^n \epsilon_p^{-1}$:

$$\mathbb{P}_e(\mathbb{P}_e(\mathbf{AB} \mid \mathbf{XY}) > \delta \text{ AND } V \geq v_{\text{thresh}}) \leq \epsilon_{\text{p}} \qquad (4)$$

where $\delta = [1 + (1 - \sqrt[n]{\epsilon_{\text{p}} v_{\text{thresh}}})/(2m)]^n$ and $\mathbb{P}_e$ denotes the probability distribution conditioned on the event $E = e$, where e is arbitrary. The expression $\mathbb{P}_e(\mathbf{AB} \mid \mathbf{XY})$ denotes the random variable that takes the value $\mathbb{P}_e(\mathbf{AB} = \mathbf{ab} \mid \mathbf{XY} = \mathbf{xy})$ when $\mathbf{ABXY}$ takes the value $\mathbf{abxy}$.

In words, this theorem says that, with high probability, if V is at least as large as v_{thresh}, then the output $\mathbf{AB}$ is unpredictable, in the sense that no individual outcome $\mathbf{AB} = \mathbf{ab}$ occurs with probability higher than δ, even given the information $\mathbf{XY}E = \mathbf{xy}e$. The theorem supports a protocol that aborts if V takes a value less than v_{thresh}, and passes otherwise. If the probability of passing were 1, then $-\log_2(\delta)$ would be a so-called 'smooth min-entropy'[6]—a quantity that characterizes the number of uniformly distributed bits of randomness that are in principle available in $\mathbf{AB}$. We show in Supplementary Information section 3 that, for constant ϵ_{p}, $-\log_2(\delta)$ is proportional to the number of trials. The number of bits that we can actually extract depends on ϵ_{fin}, the maximum allowed distance of the final output from uniform. We also show in Supplementary Information section 2 that the entropy production theorem can be proved even if the settings probabilities are not known exactly.

To extract the available randomness in $\mathbf{AB}$, we use the TMPS algorithm to obtain an extractor, specifically a function Ext that takes as inputs the string $\mathbf{AB}$ and a 'seed' bit string $\mathbf{S}$ of length d, where $\mathbf{S}$ is uniform and independent of $\mathbf{ABXY}$. Its output is a bit string of length t. $\mathbf{S}$ can be obtained from d additional instances of the random variables X_i, so equation (2) ensures the independence and uniformity conditions on $\mathbf{S}$ that are needed. For the output to be within a distance ϵ_{fin} of uniform independently of $\mathbf{XY}$ and E, the entropy production and extractor parameters must satisfy the constraints given in the 'protocol soundness theorem', which we prove in Supplementary Information section 5. In the statement of the theorem, the measure of distance used is the total variation distance, which is expressed by the left-hand side of equation (6), and 'pass' is the event that V exceeds v_{thresh}.

The protocol soundness theorem is as follows. Let $0 < \epsilon_{\text{ext}}, \kappa < 1$. Suppose that $\mathbb{P}(\text{pass}) \geq \kappa$ and that the protocol parameters satisfy

$$t + 4\log_2(t) \leq -\log_2(\delta) + \log_2(\kappa) + 5\log_2(\epsilon_{\text{ext}}) - 11 \qquad (5)$$

Then, the output $\mathbf{U} = \text{Ext}(\mathbf{AB}, \mathbf{S})$ of the function obtained by the TMPS algorithm satisfies

$$\frac{1}{2} \sum_{\mathbf{u}, \mathbf{xys}e} |\mathbb{P}(\mathbf{U} = \mathbf{u}, \mathbf{XYS}E = \mathbf{xys}e \mid \text{pass})$$
$$- \mathbb{P}^{\text{unif}}(\mathbf{U} = \mathbf{u})\mathbb{P}(\mathbf{XY}E = \mathbf{xy}e \mid \text{pass})\mathbb{P}^{\text{unif}}(\mathbf{S} = \mathbf{s})| \qquad (6)$$
$$\leq \frac{\epsilon_{\text{p}}}{\mathbb{P}(\text{pass})} + \epsilon_{\text{ext}}$$

where $\mathbb{P}^{\text{unif}}$ denotes the uniform probability distribution.

The number of seed bits d that are required satisfies $d = \mathcal{O}[\log(t)\log(nt/\epsilon_{\text{ext}})^2]$; we provide an explicit bound in Supplementary Information section 4. The protocol soundness theorem enables us to quantify the uniformity of the randomness that is produced with an overall final error parameter of $\epsilon_{\text{fin}} = \max(\epsilon_{\text{p}}/\kappa + \epsilon_{\text{ext}}, \kappa)$. (This choice of error parameter is conservative; see Supplementary Information section 5.) For any probability of passing greater than ϵ_{fin}, the total variation distance from uniform (conditionally on passing) is at most ϵ_{fin}.

We applied our protocol to five datasets using a set-up based on that described previously[13], with improvements described in Methods. Each dataset was collected in 5–10 min. Before starting the protocol, we set aside the first 5×10^6 trials of each dataset as training data, which we used to choose the parameters that are needed by the protocol. With the training data removed, the number n of trials used by the protocol was between 2.5×10^7 and 5.5×10^7 for each dataset. We used the training

data to determine a Bell function T with statistically strong violation of local realism on the training data according to the prediction-based ratio method[32]; see Supplementary Information section 3. The function T obtained for the fifth dataset, which was the longest in duration and produced the most randomness, assigned values between 0.927 and 1.004 to the 16 different experimental outcomes. We computed thresholds v_{thresh} so that a sample of n i.i.d. trials from the distribution inferred from the training data would have a high probability of exceeding v_{thresh}.

For the fifth dataset, a sample of n i.i.d. trials from the distribution inferred from the training data would have a probability of approximately 0.99 of exceeding a threshold of $v_{\text{thresh}} = 1.5 \times 10^{32}$. Exceeding this threshold would allow the extraction of 1,024 bits that are uniformly distributed to within $\epsilon_{\text{fin}} = 10^{-12}$, using $\epsilon_{\text{p}} = \kappa^2 = 9.025 \times 10^{-25}$ and $\epsilon_{\text{ext}} = 5 \times 10^{-14}$. These values were chosen on the basis of a numerical study of the constraints on the number t of bits extracted for fixed values of $\epsilon_{\text{fin}} = 10^{-12}$. Running the protocol on the remaining 55,110,210 trials with these parameters, the product $\prod_{i=1}^n T_i$ exceeded v_{thresh}, and so the protocol passed. Applying the extractor to the resulting output string $\mathbf{AB}$ with a seed of length $d = 315,844$, we extracted 1,024 bits, certified to be uniform to within 10^{-12}, the first ten of which are 1110001001. In Fig. 2 we display the extractable bits for alternative choices of ϵ_{fin} for all five datasets.

For the dataset that produced 1,024 new near random bits, our protocol used 1.10×10^8 uniform bits to choose the settings and 3.16×10^5 uniform bits to choose the seed. The strong extractor property[28] of the TMPS algorithm ensures that the seed bits are still uniform, conditional on passing, so they can be recovered at the end of the protocol for use elsewhere. This is not the case for the settings-choice bits because the probability of passing is less than 1. To reduce the entropy used for the settings, our protocol can be modified to use highly biased settings choices[5]. Reducing settings entropy is not a priority if the settings and seed bits come from a public source of randomness, in which case the output bits can still be certified to be unknown to external observers such as Eve and the current protocol is an effective method for private randomness generation[2,10].

For future work, we hope to take advantage of the adaptive capabilities of the entropy production theorem (Supplementary Information section 2) to compensate for experimental drift dynamically during
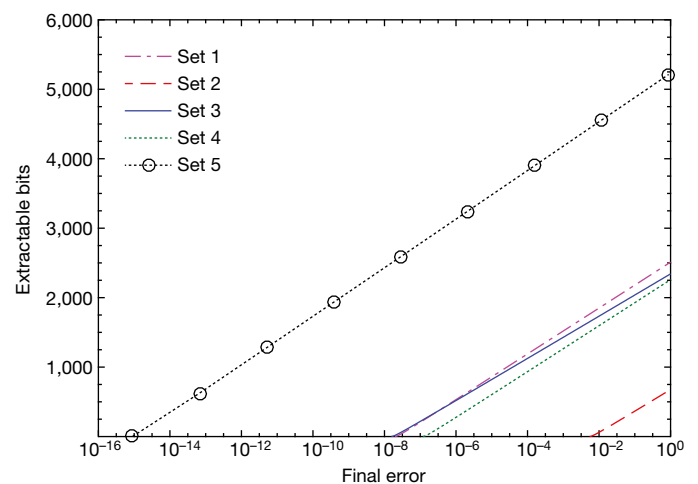

Fig. 2 | Extractable bits as a function of error. The figure shows the trade-off between the final error ϵ_{fin} and the number of extractable bits t for values of v_{thresh} pre-chosen to yield estimated passing probabilities that exceed 95%. These thresholds were met in each case. For all datasets (1–5) we set $\epsilon_{\text{p}} = \kappa^2 = (0.95\epsilon_{\text{fin}})^2$ and $\epsilon_{\text{ext}} = 0.05\epsilon_{\text{fin}}$, a split that was generally found to be near-optimal when numerically maximizing t in equation (5) for fixed values of ϵ_{fin}. The number of trials for datasets 1–5 were $n_1 = 24,865,320$, $n_2 = 24,809,970$, $n_3 = 24,818,959$, $n_4 = 24,846,822$ and $n_5 = 55,110,210$.

run time. In view of advances towards practical quantum computing, it is desirable to study the protocol when experimental devices may have long-term quantum memories and remain entangled with Eve after the protocol has begun. This may require more conservative randomness generation.

With the advent of loophole-free Bell tests, we have demonstrated that it is possible to build quantum devices that exploit quantum non-locality to remove many of the device-dependent assumptions in current technological implementations of random-number generators. Generators such as ours provide the best method currently known for physically producing randomness, thereby improving the security of a wide range of applications.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0019-0.

1. Acín, A. & Masanes, L. Certified randomness in quantum physics. *Nature* **540**, 213–219 (2016).
2. Pironio, S. & Massar, S. Security of practical private randomness generation. *Phys. Rev. A* **87**, 012336 (2013).
3. Miller, C. A. & Shi, Y. Robust protocols for securely expanding randomness and distributing keys using untrusted quantum devices. *J. ACM* **63**, 33 (2016).
4. Colbeck, R. & Kent, A. Private randomness expansion with untrusted devices. *J. Phys. A* **44**, 095305 (2011).
5. Pironio, S. et al. Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).
6. Vazirani, U. & Vidick, T. Certifiable quantum dice - or, exponential randomness expansion. In *STOC'12 Proc. 44th Annual ACM Symposium on Theory of Computing* (ed. Pitassi, T.) 61–76 (2012).
7. Fehr, S., Gelles, R. & Schaffner, C. Security and composability of randomness expansion from Bell inequalities. *Phys. Rev. A* **87**, 012335 (2013).
8. Chung, K.-M., Shi, Y. & Wu, X. Physical randomness extractors: generating random numbers with minimal assumptions. Preprint at https://arxiv.org/abs/1402.4797 (2014).
9. Nieto-Silleras, O., Pironio, S. & Silman, J. Using complete measurement statistics for optimal device-independent randomness evaluation. *New J. Phys.* **16**, 013035 (2014).
10. Bancal, J.-D., Sheridan, L. & Scarani, V. More randomness from the same data. *New J. Phys.* **16**, 033011 (2014).
11. Thinh, L., de la Torre, G., Bancal, J.-D., Pironio, P. & Scarani, V. Randomness in post-selected events. *New J. Phys.* **18**, 035007 (2016).
12. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
13. Shalm, L. K. et al. Strong loophole-free test of local realism. *Phys. Rev. Lett.* **115**, 250402 (2015).
14. Giustina, M. et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Phys. Rev. Lett.* **115**, 250401 (2015).
15. Paar, C. & Pelzl, J. *Understanding Cryptography* (Springer, Heidelberg, 2010).
16. Fischer, M. J., Iorga, M. & Peralta, R. A public randomness service. In *Proc. International Conference on Security and Cryptography (SECRYPT 2011)* (eds Lopez, J. & Samarati, P.) 434–438 (2011).
17. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Phys. Phys. Fiz.* **1**, 195–200 (1964).
18. Bell, J. S., Shimony, A., Horne, M. A. & Clauser, J. F. An exchange on local beables. *Dialectica* **39**, 85–96 (1985).
19. Mayers, D. & Yao, A. Quantum cryptography with imperfect apparatus. In *FOCS'98 Proc. 39th Annual Symposium on Foundations of Computer Science* (ed Motwani, R.) 503–509 (1998).
20. Acín, A. et al. Device-independent security of quantum cryptography against collective attacks. *Phys. Rev. Lett.* **98**, 230501 (2007).
21. Liu, Y. et al. High-speed device-independent quantum random number generation without a detection loophole. *Phys. Rev. Lett.* **120**, 010503 (2018).
22. Rosenfeld, W. et al. Event-ready Bell test using entangled atoms simultaneously closing detection and locality loopholes. *Phys. Rev. Lett.* **119**, 010402 (2017).
23. Abellán, C. et al. Challenging local realism with human choices. *Nature* (in the press).
24. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Rev. Mod. Phys.* **86**, 419–478 (2014).
25. Cirel'son, B. S. Quantum generalizations of Bell's inequality. *Lett. Math. Phys.* **4**, 93–100 (1980).
26. Navascués, M., Pironio, S. & Acín, A. A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations. *New J. Phys.* **10**, 073013 (2008).
27. Trevisan, L. Extractors and pseudorandom generators. *J. Assoc. Comput. Mach.* **48**, 860–879 (2001).
28. Mauerer, W., Portmann, C. & Scholz, V. B. A modular framework for randomness extraction based on Trevisan's construction. Preprint at https://arxiv.org/abs/1212.0520 (2012).
29. Coudron, M. & Yuen, H. Infinite randomness expansion with a constant number of devices. In *STOC'14 Proc. 46th Annual ACM Symposium on Theory of Computing* (ed Shmoys, D.) 427–436 (2014).
30. Arnon-Friedman, R., Dupuis, F., Fawzi, O., Renner, R. & Vidick, T. Practical device-independent quantum cryptography via entropy accumulation. *Nat. Commun.* **9**, 459 (2018).
31. Miller, C. & Shi, Y. Universal security for randomness expansion from the spot-checking protocol. *SIAM J. Comput.* **46**, 1304–1335 (2017).
32. Zhang, Y., Glancy, S. & Knill, E. Asymptotically optimal data analysis for rejecting local realism. *Phys. Rev. A* **84**, 062118 (2011).

METHODS

We used polarization-entangled photons generated by a nonlinear crystal pumped by a pulsed, picosecond laser at approximately 775 nm in a configuration similar to that reported previously[13], but with several improvements to increase the rate of randomness extraction. The repetition rate of the laser was 79.3 MHz and each pulse that entered the crystal had a probability of approximately 0.003 of creating an entangled photon pair in the state $|\psi\rangle \approx 0.982|HH\rangle + 0.191|VV\rangle$ at a centre wavelength of 1,550 nm. By pumping the crystal with approximately five times as much power, and using a 20-mm-long crystal, we were able to increase the per-pulse probability of generating a down-conversion event substantially compared with the previous configuration[13] while maintaining similar overall system efficiencies. The two entangled photons from each pair were sent separately to one of the two measurement stations, which were 187 ± 1 m apart. At Alice and Bob, a Pockels cell and a polarizer combined to allow the rapid switching of measurement bases and the measurement of the polarization state of the incoming photons. Alice's computed optimal polarization measurement angles, relative to a vertical polarizer, were $a = -3.7°$ and $a' = 23.6°$, and Bob's were $b = 3.7°$ and $b' = -23.6°$. Each Pockels cell operated at a rate of 100 kHz, allowing us to perform 100,000 trials per second (the driver electronics on the Pockels cells sets this rate). A 10-MHz oscillator kept Alice's and Bob's time-tagger clocks locked. After passing through the polarization optics, the photons were each coupled into a single-mode fibre and detected using superconducting single-photon nanowire detectors, with Bob's detector operating at approximately 90% efficiency and Alice's detector operating with approximately 92% efficiency[33]. For this experiment, the total symmetric system heralding efficiency was $75.5\% \pm 0.5\%$, which is greater than the 71.5% threshold that is required to close the detection loophole for our experimental configuration after accounting for unwanted background counts at our detectors and slight imperfections in our state-preparation and measurement components.

With this configuration, Bob completed his measurement 294.4 ± 3.7 ns before a hypothetical switching signal travelling at light speed from Alice's Pockels cell could arrive at his station. Similarly, Alice completed her measurement 424.2 ± 3.7 ns before such a signal from Bob's Pockels cell could arrive at her location. The outcome values for each trial were obtained by aggregating the photon detection or non-detection events from several short time intervals, each lasting 1,024 ps and timed to correspond to one pulse of the pump laser. If any photons were detected in the short intervals, then the outcome was '+'; if no photons were detected, then the outcome was '0'. The previous experiment[13] used at most 7 short intervals, but here we were able to include 14 intervals while maintaining space-like separation, which further increased the probability of observing a photon during each trial. For demonstration purposes, Alice and Bob each used Python's random.py module with the default generator (the Mersenne twister) to pick their settings at each trial. This pseudorandom source is predictable, and for secure applications of the protocol in an adversarial scenario, such as if the photon pair source or measurement devices are obtained from an untrusted provider, settings choices must be based on random sources that are effectively not predictable. However, from our knowledge of device construction, we know that our devices have no physical resources for predicting pseudorandom numbers and expect that the measurement settings were effectively independent of the relevant devices so that equations (2) and (3) still hold. We remark that the settings choices for the previous datasets[13] were based on physical random sources.

With the improved detection efficiency, the higher per-trial probability for Alice and Bob to detect a photon, and a higher signal-to-background counts ratio, we are able to improve the magnitude of our Bell violation and to reduce the number of trials that are required to achieve a statistically significant violation by an order of magnitude.

Sample size. No statistical methods were used to predetermine sample size.

Data availability. The photon detection data that support the findings of this study are available in the NIST Published Data Repository (https://doi.org/10.18434/T4/1423448).

33. Marsili, F. et al. Detecting single infrared photons with 93% system efficiency. *Nat. Photon.* **7**, 210–214 (2013).

LETTER

https://doi.org/10.1038/s41586-018-0007-4

Anomalously weak Labrador Sea convection and Atlantic overturning during the past 150 years

David J. R. Thornalley[1,2]*, Delia W. Oppo[2], Pablo Ortega[3], Jon I. Robson[3], Chris M. Brierley[1], Renee Davis[1], Ian R. Hall[4], Paola Moffa-Sanchez[4], Neil L. Rose[1], Peter T. Spooner[1], Igor Yashayaev[5] & Lloyd D. Keigwin[2]

The Atlantic meridional overturning circulation (AMOC) is a system of ocean currents that has an essential role in Earth's climate, redistributing heat and influencing the carbon cycle[1,2]. The AMOC has been shown to be weakening in recent years[1]; this decline may reflect decadal-scale variability in convection in the Labrador Sea, but short observational datasets preclude a longer-term perspective on the modern state and variability of Labrador Sea convection and the AMOC[1,3–5]. Here we provide several lines of palaeo-oceanographic evidence that Labrador Sea deep convection and the AMOC have been anomalously weak over the past 150 years or so (since the end of the Little Ice Age, LIA, approximately AD 1850) compared with the preceding 1,500 years. Our palaeoclimate reconstructions indicate that the transition occurred either as a predominantly abrupt shift towards the end of the LIA, or as a more gradual, continued decline over the past 150 years; this ambiguity probably arises from non-AMOC influences on the various proxies or from the different sensitivities of these proxies to individual components of the AMOC. We suggest that enhanced freshwater fluxes from the Arctic and Nordic seas towards the end of the LIA—sourced from melting glaciers and thickened sea ice that developed earlier in the LIA—weakened Labrador Sea convection and the AMOC. The lack of a subsequent recovery may have resulted from hysteresis or from twentieth-century melting of the Greenland Ice Sheet[6]. Our results suggest that recent decadal variability in Labrador Sea convection and the AMOC has occurred during an atypical, weak background state. Future work should aim to constrain the roles of internal climate variability and early anthropogenic forcing in the AMOC weakening described here.

The AMOC comprises northward transport of warm surface and thermocline waters, and their deep southward return flow as dense waters that formed through cooling processes and sinking at high latitudes[2]. The stability of the AMOC in response to ongoing and projected climate change is uncertain. Monitoring of the AMOC during the past decade with an instrument array at 26° N has suggested that the AMOC is weakening, and that this is occurring ten times faster than would be expected from climate model projections[1]. However, it remains uncertain whether this trend is part of a longer-term decline, natural multidecadal variability, or a combination of both. Here, we develop past reconstructions of AMOC variability that can be compared directly with instrumental datasets and provide longer-term perspective.

The Labrador Sea is an important region for deep-water formation in the North Atlantic ocean[5]. Moreover, modelling studies suggest that deep-Labrador-Sea density (DLSD) might be a useful predictor of AMOC change[3,4,7]. This is because density anomalies produced in the Labrador Sea—caused predominantly by varying deep convection—can propagate southwards rapidly (on timescales of the order of months) along the western margin via boundary waves, altering the cross-basin zonal density gradient, and thus modifying geostrophic transport and therefore AMOC strength[2–4,7–9]. Building upon these studies, we show that DLSD anomalies are also associated with changes in the velocity of the deep western boundary current (DWBC) and the strength of the AMOC at 45° N in the high-resolution climate model HadGEM3-GC2 (see Methods and Fig. 1).

In addition to this link between the AMOC, DLSD and DWBC, changes in the AMOC also alter ocean heat transport. Modelling studies[10] suggest that AMOC weakening affects the upper-ocean heat content of the subpolar gyre (SPG) with a lag time of around ten years. Moreover, a distinct AMOC fingerprint on subsurface temperatures (T_{sub}; at depths of 400 m)[11] characterizes weak AMOC phases, with a dipole pattern of warming of the Gulf Stream extension region[12] and cooling of the subpolar Northeast Atlantic. We exploit here the model-based covariance of decadal changes in the AMOC with DLSD anomalies, SPG upper-ocean heat content and the T_{sub} fingerprint, to extend constraints on past AMOC variability (see Methods). Over the instrumental era (from AD 1950 or so), these indices suggest substantial decadal variability in the AMOC, with coherent changes in DLSD, lagged SPG upper-ocean heat content and a lagged T_{sub} AMOC fingerprint[3,5,8,10,11].

The model results in Fig. 1 suggest that we can use flow-speed reconstructions of the DWBC to infer past changes in DLSD and the AMOC. We analysed the sortable-silt mean grain size—a proxy for near-bottom current flow speed[13]—in two marine sediment cores (48JPC and 56JPC; see Methods and Extended Data Figs. 1, 2) located under the influence of southward-flowing Labrador Sea Water (LSW) within the DWBC off Cape Hatteras (hereafter $DWBC_{LSW}$). The high sediment-accumulation rates (about 0.5–1 cm per year) and the modern core-top enable direct comparison of the record from 56JPC with observational datasets (Fig. 2).

In agreement with the model-predicted relationship (Fig. 1), changes in the inferred flow speed of the $DWBC_{LSW}$ show similar, in-phase, variability with observed DLSD[5]. Moreover, there is strong covariability of our $DWBC_{LSW}$ proxy with the lagged (12-year) SPG upper-ocean heat content and T_{sub} index from observational analysis (Fig. 2a). Over the past 100 years or so, the spatial correlation of upper-ocean heat content anomalies associated with our $DWBC_{LSW}$ proxy has closely resembled the T_{sub} AMOC fingerprint (Fig. 2b, c), supporting the concept that the $DWBC_{LSW}$ proxy and upper-ocean temperature changes provide complementary, coherent information on a common phenomenon, namely, AMOC variability. Combined, these datasets indicate that decadal variability has been a dominant feature of the past 130 years, with the most recent strengthening of LSW formation during the mid-1990s, and its subsequent decline, being particularly prominent features.

To gain insight into variability before the instrumental era, we first extended our $DWBC_{LSW}$ flow-speed reconstruction (Fig. 3e). The $DWBC_{LSW}$ proxy suggests that the AMOC has been weaker during the past 150 years than at any other time during the past 1,600 years. The emergence of this weaker state (during which the smoothed record exceeds a noise threshold of 2σ pre-industrial-era variability) takes place at about AD 1880 in both cores. The overall transition

---

[1]Department of Geography, University College London, London, UK. [2]Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA, USA. [3]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK. [4]School of Earth and Ocean Sciences, Cardiff University, Cardiff, UK. [5]Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada. *e-mail: d.thornalley@cantab.net
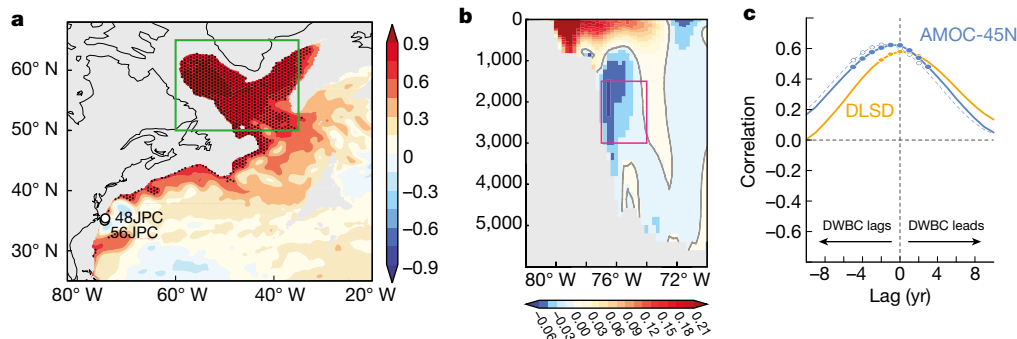
Fig. 1 | Modelled link between DWBC velocity, deep Labrador Sea density and the AMOC. a, Correlation (colour bar on the right) between the vertically averaged ocean density (1,000 m to 2,500 m) and DLSD (average ocean density between 1,000 m and 2,500 m in the area defined by the green box), as modelled using a control run of the high-resolution climate model HadGEM3-GC2. The locations of the sediment-core sites used for DWBC flow-speed reconstruction are also shown. **b**, Climatology of the modelled meridional ocean velocity (in metres per second; see colour bar) at 30° N to 35° N (see Methods and Extended Data Figs. 7, 8), illustrating the modelled position of the DWBC (red outline). The y axis shows water depth in metres. **c**, Cross-correlations between the modelled average DWBC flow speed from the red box in panel **b**, and indices of DLSD and the AMOC at 45° N (the dashed blue line omits the Ekman component).

occurs from about AD 1750 to AD 1900—late in the Little Ice Age (AD 1350–1850) and during the early stages of the industrial era (1830 onwards[14]). Applying the flow-speed calibration for sortable silt[13] suggests a decrease from $17 \, \text{cm s}^{-1}$ to $14.5 \, \text{cm s}^{-1}$ in core 56JPC during this transition period, and from $14 \, \text{cm s}^{-1}$ to $12 \, \text{cm s}^{-1}$ in 48JPC, suggesting a decrease in DWBC_{LSW} strength of approximately 15% (assuming a constant DWBC_{LSW} cross-sectional area). This decrease is equivalent to 3σ and 4σ of the pre-industrial-era variability in 48JPC and 56JPC, respectively.

Second, we compiled quantitative proxy records of subsurface ocean temperatures (at depths of about 50–200 m) from key locations to extend the T_{sub} AMOC proxy (Fig. 3a–c; see Methods and Extended Data Figs. 3, 4). This T_{sub} proxy reconstruction provides support for the proposed AMOC weakening. Opposing temperature anomalies recorded in the two regions after about AD 1830—with warming of the Gulf Stream extension region and cooling of the subpolar Northeast Atlantic region—suggest a weaker industrial-era AMOC. Further support for the AMOC weakening is suggested by the spatial pattern of T_{sub} change in the Northwest Atlantic during the onset of the industrial era (Extended Data Fig. 5). In contrast to the prominent changes recorded in our proxy reconstructions at the end of the LIA, more subdued variability occurs during the earlier part of our records (AD 400–1800). This might suggest that the forcing and AMOC response were weaker then, or that the AMOC did not play a leading role in the (multi)centennial climate variability of this period[15,16].

Labrador Sea deep convection is a major contributor to the AMOC, but susceptible to weakening[5]. This fact, combined with its role in decadal AMOC variability over the past 100 years or so (Fig. 2) and model analysis of mechanisms for AMOC variability in operation today[8], makes it likely that changes in Labrador Sea convection were involved in the weakening of the AMOC at the end of the LIA. Further correlative (although not necessarily causative) support for this idea is revealed by palaeo-oceanographic evidence from the Labrador Sea. Strong deep convection in the Labrador Sea is typically associated with cooling and freshening of the subsurface ocean[5]. Therefore, the reconstructed shift to warmer and saltier subsurface conditions in the northeast Labrador Sea[17] over the past 150 years (Fig. 3d; equivalent to around 2σ of pre-industrial-era variability) is consistent with a shift to a state characterized by reduced deep convection, with only occasional episodes of sustained deep convection. Reconstructions of the
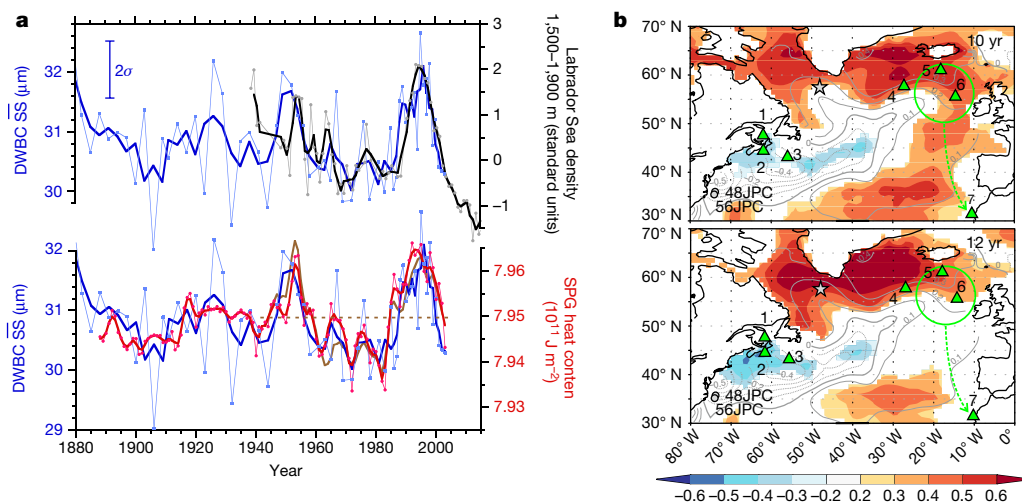

Fig. 2 | Proxy validation and recent multidecadal variability. a, The mean grain size of sortable silt (SS; from sediment core 56JPC; blue) is compared with: the central-Labrador-Sea annual density[5] (black; $r^2 = 0.56$; $n = 54$), which is comparable to the model-based DLSD (Extended Data Fig. 9); the 12-year lagged SPG upper-ocean heat content (at 0–700 m; 55° N to 65° N, 15° W to 60° W; EN4 dataset; red; $r^2 = 0.58$; $n = 116$); and the 12-year lagged T_{sub} AMOC fingerprint[11] (brown; dashed line shows the zero line; $r^2 = 0.76$; $n = 55$). Correlations (and the 2σ SS error bar; $n = 30$) are for three-point means (thicker lines). Low-resolution 48JPC data are not shown. **b**, 10- and 12-year lagged spatial correlation (colour bar; R) of upper-ocean heat content (at 0–700 m) with reconstructed DWBC_{LSW} flow speed (from sediment core 56JPC); the heat content lags behind the DWBC. Grey contours show the spatial T_{sub} AMOC proxy[11]; green triangles show T_{sub} proxy sites; the green circle marks the surface region controlling benthic temperatures at site 7; grey circles are DWBC sites; the grey star marks the core site from ref. [17].

other major deep-water contributors to the AMOC—the two Nordic Seas overflows—suggest that, on centennial timescales, they have varied in anti-phase and probably therefore compensated for one another during the past 3,000 years[18]. Hence, changes in Labrador Sea deep convection may have been the main cause of AMOC variability over this period.

Although atmospheric circulation has played a dominant part in recent decadal variability in the AMOC (and LSW)[2,8], there is no strong evidence that the AMOC decrease at the end of the LIA was similarly caused by a shift in atmospheric circulation[19]. Instead, we hypothesize that the AMOC weakening was caused by enhanced freshwater fluxes associated with the melting and export of ice and freshwater from the Arctic and Nordic seas. During the LIA, circum-Arctic glaciers and multiyear Arctic and Nordic sea ice were at their most advanced state of the past few thousand years, and there were large ice shelves in the Canadian Arctic and exceptionally thick multiyear sea ice. Yet, by the early twentieth century, many of these features had disappeared or were retreating[20-23].

Modelling studies suggest that enhanced freshwater fluxes of about 10–100 mSv over a few decades can weaken Labrador Sea convection and the AMOC[24], although models with strong hysteresis of Labrador Sea convection[25] suggest that this weakening may be caused by as little as 5–10 mSv of freshwater. Unfortunately, there are few data to constrain the Arctic and Nordic Sea freshwater fluxes associated with the end of the LIA. The earliest observational datasets[26,27] suggest that a flux of about 10 mSv resulted from sea-ice loss in the Arctic and Nordic seas during 1895–1920, to which we must also add melting of previously expanded circum-Arctic glaciers and ice shelves, and enhanced melting of the Greenland Ice Sheet. Alternatively, we could estimate that a 1-m reduction in average Arctic sea-ice thickness during the termination of the LIA could have yielded a freshwater flux of 10 mSv for 50 years. Although further work is required to improve this incomplete estimate, there was probably sufficient freshwater stored in the Arctic and Nordic seas during the LIA to influence Labrador Sea convection and the AMOC.

The AMOC weakening recorded in our two marine reconstructions is broadly similar to that in a predominantly terrestrial-based AMOC proxy reconstruction[6] (Fig. 3c). Our T_{sub} AMOC proxy and that of ref. [6] (Fig. 3c) both suggest a substantial decline in the AMOC through the twentieth century, whereas our $DWBC_{LSW}$ AMOC proxy and the observational-based T_{sub} AMOC index (Fig. 2a and Extended Data Fig. 6) suggest relatively little long-term AMOC decline during this period. These differences may be attributed to several factors. First, our sediment-core-based T_{sub} proxy is subject to artificial smoothing, caused by combining numerous records with substantial (around 10–100-year) individual age uncertainties, and compounded by sediment mixing by organisms (bioturbation). Furthermore, the T_{sub} proxy sediment cores were retrieved in the late 1990s and early 2000s, and so cannot capture the strong T_{sub} index recovery from around 2000 to 2010 that reverses the earlier prolonged decline (Extended Data Fig. 6). Alternatively, the earlier, more threshold-like change in the $DWBC_{LSW}$ AMOC proxy may be due to local shifts in the position of the DWBC and/or nonlinear dynamics of the DWBC response to AMOC change. However, given the similarity of the $DWBC_{LSW}$ reconstructions from cores 56JPC and 48JPC (located at different water depths), and the strong correlation of $DWBC_{LSW}$ with Labrador Sea density and the T_{sub} AMOC index over the instrumental period, we suggest that these factors are not substantial. Finally, the differences between the AMOC reconstructions may reflect their varying response timescales and sensitivities to the different individual components of the AMOC and the SPG[28,29].

Our study raises several issues regarding the modelling of the AMOC in historical experiments. The inferred transition to a weakened AMOC occurred near the onset of the industrial era, several decades before the strongest global warming trend, and has remained weak up to the present day. This suggests either hysteresis of the AMOC in response to an early climate forcing—natural (solar, volcanic) or anthropogenic (greenhouse gases, aerosols, land-use change)—or that continued climate forcing, such as the melting of the Greenland Ice Sheet[6], has been sufficient to keep the AMOC weak or cause further weakening.
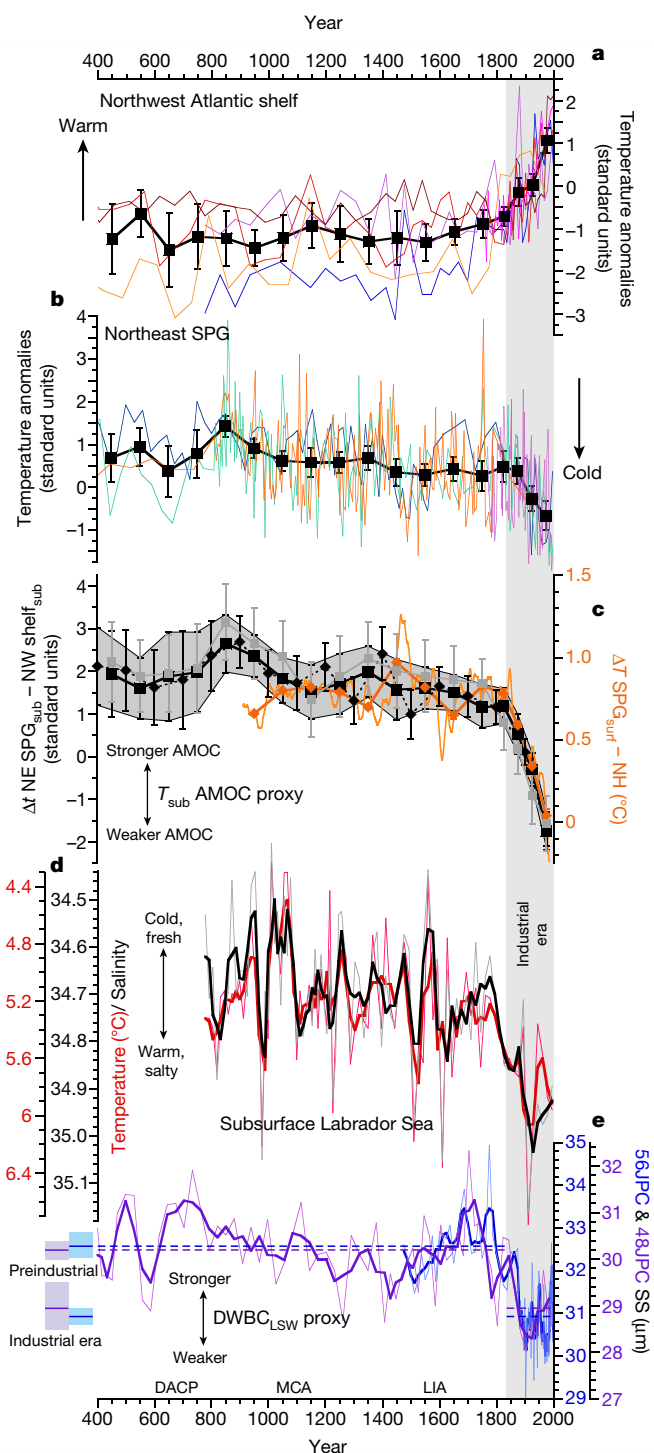

Fig. 3 | Proxy reconstructions of AMOC changes over the past 1,600 years. a, b, Subsurface Northwest Atlantic shelf (**a**) and Northeast Atlantic SPG (**b**) temperatures, taken at sites shown in Fig. 2b. Composite stacks are in black. **c**, Black and grey, our T_{sub} AMOC proxy with different types of binning (see Extended Data Fig. 4). Orange: AMOC proxy from Rahmstorf et al.[6]; 1 °C = 2.3 Sv; thin line, 21-year smoothing; thick line and symbols, binned as for our T_{sub} AMOC proxy. NE SPG, Northeast Atlantic subpolar gyre; NW shelf, Northwest Atlantic shelf; NH, Northern Hemisphere; sub, subsurface; surf, surface. **d**, Subsurface (around 100–200 m) temperature and salinity of the northeast Labrador Sea, based on Mg/Ca-δ^{18}O analysis of the planktic foraminifera *Neogloboquadrina pachyderma*[17]. **e**, Sortable silt (SS) mean grain size. Blue, core 56JPC; purple, 48JPC; bold, three-point means; dashed lines, industrial/pre-industrial era averages; error bars/shading, ±2 s.e. DACP, Dark Ages Cold Period (around AD 400–800); MCA, Mediaeval Climate Anomaly (around AD 900–1250).

Our reconstructions also differ from most climate model simulations, which show either negligible AMOC change or a later, more gradual reduction[30]. Many factors may be responsible for this model–data discrepancy: a misrepresentation of AMOC-related processes and possible hysteresis, including underestimation of AMOC sensitivity to climate (freshwater) forcing[29,31]; the underestimation or absence of important freshwater fluxes during the end of the LIA; and the lack of transient forced behaviour in the 'constant forcing' pre-industrial controls used to initialize historical forcings. Resolving these issues will be important for improving the accuracy of projected changes in the AMOC.

In conclusion, our study reveals an anomalously weak AMOC over the past 150 years or so. Because of its role in heat transport, it is often assumed that AMOC weakening cools the Northern Hemisphere. However, our study demonstrates that changes in the AMOC are not always synchronous with temperature changes. That AMOC weakening occurred during the late LIA and onset of the industrial era, rather than earlier in the LIA, may point to additional forcing factors at this time, such as an increase in the export of thickened Arctic and Nordic sea ice, or the melting of circum-Arctic ice shelves. The persistence of a weak AMOC during the twentieth century, when there was pronounced Northern Hemisphere and global warming, suggests that other climate forcings—such as greenhouse gas warming—were dominant during this period. We therefore infer that the AMOC has responded to recent centennial-scale climate change, rather than driven it. Regardless, the weak state of the AMOC over the past 150 years may have modified northward ocean heat transport, as well as atmospheric warming by altering ocean–atmosphere heat transfer[32,33], underscoring the need for continued investigation of the role of the AMOC in climate change. Determining the future behaviour of the AMOC will depend in part on constraining its sensitivity and possible hysteresis to freshwater input, for which improved historical estimates of these fluxes during the AMOC weakening reported here will be especially useful.

Online content

1. Srokosz, M. A. & Bryden, H. L. Observing the Atlantic Meridional Overturning Circulation yields a decade of inevitable surprises. *Science* **348**, 1255575 (2015).
2. Buckley, M. W. & Marshall, J. Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: a review. *Rev. Geophys.* **54**, 5–63 (2016).
3. Jackson, L. C., Peterson, K. A., Roberts, C. D. & Wood, R. A. Recent slowing of Atlantic overturning circulation as a recovery from earlier strengthening. *Nat. Geosci.* **9**, 518–522 (2016).
4. Robson, J., Hodson, D., Hawkins, E. & Sutton, R. Atlantic overturning in decline? *Nat. Geosci.* **7**, 2–3 (2014).
5. Yashayaev, I. Hydrographic changes in the Labrador Sea, 1960–2005. *Prog. Oceanogr.* **73**, 242–276 (2007).
6. Rahmstorf, S. et al. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Change* **5**, 475–480 (2015); corrigendum **5**, 596 (2015).
7. Hodson, D. L. R. & Sutton, R. T. The impact of resolution on the adjustment and decadal variability of the Atlantic meridional overturning circulation in a coupled climate model. *Clim. Dyn.* **39**, 3057–3073 (2012).
8. Ortega, P., Robson, J., Sutton, R. T. & Andrews, M. B. Mechanisms of decadal variability in the Labrador Sea and the wider North Atlantic in a high-resolution climate model. *Clim. Dyn.* **49**, 2625–2647 (2016).
9. Roberts, C. D., Garry, F. K. & Jackson, L. C. A multimodel study of sea surface temperature and subsurface density fingerprints of the Atlantic Meridional Overturning Circulation. *J. Clim.* **26**, 9155–9174 (2013).
10. Robson, J., Ortega, P. & Sutton, R. A reversal of climatic trends in the North Atlantic since 2005. *Nat. Geosci.* **9**, 513–517 (2016).
11. Zhang, R. Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation. *Geophys. Res. Lett.* **35**, L20705 (2008).
12. Saba, V. S. et al. Enhanced warming of the Northwest Atlantic Ocean under climate change. *J. Geophys. Res. Oceans* **121**, 118–132 (2016).
13. McCave, I. N., Thornalley, D. J. R. & Hall, I. R. Relation of sortable silt grain-size to deep-sea current speeds: calibration of the 'mud current meter'. *Deep Sea Res. I Oceanogr. Res. Pap.* **127**, 1–12 (2017).
14. Abram, N. J. et al. Early onset of industrial-era warming across the oceans and continents. *Nature* **536**, 411–418 (2016).
15. Moreno-Chamarro, E., Zanchettin, D., Lohmann, K. & Jungclaus, J. H. An abrupt weakening of the subpolar gyre as trigger of Little Ice Age-type episodes. *Clim. Dyn.* **48**, 727–744 (2017).
16. Miller, G. H. et al. Abrupt onset of the Little Ice Age triggered by volcanism and sustained by sea-ice/ocean feedbacks. *Geophys. Res. Lett.* **39**, L02708 (2012).
17. Moffa-Sánchez, P., Hall, I. R., Barker, S., Thornalley, D. J. R. & Yashayaev, I. Surface changes in the eastern Labrador Sea around the onset of the Little Ice Age. *Paleoceanography* **29**, 160–175 (2014).
18. Moffa-Sanchez, P., Hall, I. R., Thornalley, D. J. R., Barker, S. & Stewart, C. Changes in the strength of the Nordic Seas overflows over the past 3000 years. *Quat. Sci. Rev.* **123**, 134–143 (2015).
19. Ortega, P. et al. A model-tested North Atlantic Oscillation reconstruction for the past millennium. *Nature* **523**, 71–74 (2015).
20. Bradley, R. S. & England, J. H. The Younger Dryas and the sea of ancient ice. *Quat. Res* **70**, 1–10 (2008).
21. Funder, S. et al. A 10,000-year record of Arctic Ocean sea-ice variability—view from the beach. *Science* **333**, 747–750 (2011).
22. Vincent, W. F., Gibson, J. A. E. & Jeffries, M. O. Ice-shelf collapse, climate change, and habitat loss in the Canadian high Arctic. *Polar Rec.* **37**, 133–142 (2001).
23. Cabedo-Sanz, P., Belt, S. T., Jennings, A. E., Andrews, J. T. & Geirsdóttir, Á. Variability in drift ice export from the Arctic Ocean to the North Icelandic Shelf over the last 8000 years: a multi-proxy evaluation. *Quat. Sci. Rev.* **146**, 99–115 (2016).
24. Yang, Q. et al. Recent increases in Arctic freshwater flux affects Labrador Sea convection and Atlantic overturning circulation. *Nat. Comm* **7**, 10525 (2016).
25. Schulz, M., Prange, M. & Klocker, A. Low-frequency oscillations of the Atlantic Ocean meridional overturning circulation in a coupled climate model. *Clim. Past* **3**, 97–107 (2007).
26. Polyakov, I. V. et al. Arctic Ocean freshwater changes over the past 100 years and their causes. *J. Clim.* **21**, 364–384 (2008).
27. Vinje, T. Anomalies and trends of sea-ice extent and atmospheric circulation in the Nordic seas during the period 1864–1998. *J. Clim.* **14**, 255–267 (2001).
28. Drijfhout, S., Oldenborgh, G. J. V. & Cimatoribus, A. Is a decline of AMOC causing the warming hole above the North Atlantic in observed and modeled warming patterns? *J. Clim.* **25**, 8373–8379 (2012).
29. Sgubin, G., Swingedouw, D., Drijfhout, S., Mary, Y. & Bennabi, A. Abrupt cooling over the North Atlantic in modern climate models. *Nat. Comm* **8**, 14375 (2017).
30. Weaver, A. J. et al. Stability of the Atlantic meridional overturning circulation: a model intercomparison. *Geophys. Res. Lett.* **39**, L20709 (2012).
31. Liu, W., Xie, S.-P., Liu, Z. & Zhu, J. Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate. *Sci. Adv.* **3**, e1601666 (2017).
32. Drijfhout, S. Competition between global warming and an abrupt collapse of the AMOC in Earth's energy imbalance. *Sci. Rep.* **5**, 14877 (2015).
33. Kostov, Y., Armour, K. C. & Marshall, J. Impact of the Atlantic meridional overturning circulation on ocean heat storage and transient climate change. *Geophys. Res. Lett.* **41**, 2108–2116 (2014).

METHODS

Climate model investigation of AMOC and DWBC changes. The climate model used here was the UK Met Office's Global Coupled model 2.0 (HadGEM3–GC2). The ocean model for HadGEM3-GC2 is Global Ocean version 5.0, which is based on version 3.4 of the Nucleus for European Models of the Ocean model (NEMO)[34]. The ocean model has 75 vertical levels and is run at a nominal 1/4° resolution using the NEMO tri-polar grid. The atmospheric component is Global Atmosphere version 6.0 of the UK Met Office Unified Model, and is run at N216 resolution (around 60 km in mid-latitudes), with 85 vertical levels. More information about the model can be found in ref. [35]. The experiment analysed here was a 310-year control simulation of HadGEM3-GC2—that is, it includes no changes in external forcings. This experiment was previously run and analysed in ref. [8], where details of the specific model experiment are included. This coupled simulation has a relatively high spatial resolution for a more accurate representation of the boundary currents, and is sufficiently long to resolve a large number of decadal oscillations. All model data have been linearly detrended to remove any potential drift, and smoothed with a 10-year running mean in order to focus on the decadal and multidecadal variability.

We use the model-based relationships to support our interpretation of the proxy-based AMOC reconstructions, which cannot be validated with the limited observations available. We chose the AMOC at 45° N because this is the latitude with the largest correlations with both the DLSD and the DWBC velocity index in the model. AMOC indices defined at other latitudes (for example, 35° N or 40° N) produce weaker, but still substantial, correlations with both DLSD and the DWBC. The simulated DWBC velocity index is the average of that at 30° N to 35° N, because at 35° N (where the sediment cores were taken) the DWBC is found offshore, which we believe is associated with the model's Gulf Stream separating further north than in the observations (Extended Data Fig. 7). We note, however, that changes in the position of the observed Gulf Stream do not appear to directly control the reconstructed flow-speed changes in the DWBC$_{LSW}$ (Extended Data Fig. 10).

We have also assessed the robustness of the model-based relationships to the smoothing. For example, we reproduced the cross-correlation analysis in Fig. 1c using undetrended and/or unsmoothed data instead. In all cases, the lead–lag relationships are similar, with larger correlations emerging when the decadal smoothing is applied. Furthermore, we also tested the sensitivity of the model-based relationships to the specific model used. In particular, we repeated the analysis of Fig. 1 in the 340-year control experiment using the HiGEM climate model[36]. HiGEM has a similar horizontal ocean resolution (1/3°), but is based on a different ocean model. Encouragingly, Extended Data Fig. 8 shows that the results are consistent across the two models, in particular the link between DLSD and the DWBC, and between the DWBC and the AMOC at 45° N. However, there are some caveats. For example, both models' Gulf Streams separate too far north, which led us to define the DWBC flow indices slightly south of the core sites. HiGEM also has a deeper DWBC than that of HadGEM3-GC2. Therefore, the DWBC index was computed at different levels in both models to represent the link between DLSD and the DWBCs. However, despite these differences, both models support the general interpretation that the DWBC in the vicinity of Cape Hatteras is strongly connected with changes in the DLSD and the AMOC.

The interpretation of the model results is consistent with previously published model studies (both low and high resolution) that have revealed a coupling between the AMOC and/or Labrador Sea density, and the DWBC[3,7,11,37]. These modelled relationships support a causal link for the correlations between the instrumental records of Labrador Sea density and the reconstructed DWBC velocity, presented in Fig. 2. Furthermore, recent instrumental data for the DWBC at 39° N from 2004 to 2014 reveal that a reduction in the velocity of classical LSW within the DWBC is also accompanied by a decrease in its density[38], as hypothesized here. The observed decrease in the velocity and density of classical LSW within the DWBC between 2004 and 2014 is also consistent with the decrease in the DLSD over this period (Fig. 2a and Extended Data Fig. 9), although a longer observational DWBC time series is needed to gain confidence in this relationship.

Age models. New and updated age models for the cores are presented in Extended Data Figs. 1 and 2, and are based on ^{14}C, ^{210}Pb and spheroidal carbonaceous particle (SCP) concentration profiles[39].

Sortable silt data. We used two marine sediment cores for DWBC flow-speed reconstruction: KNR-178-56JPC (at 35° 28′ N, 74° 43′ W, 1,718 m water depth) and KNR-178-48JPC (35° 46′ N, 74° 27′ W, 2,009 m water depth). Sediments were processed using established methods[40], taking 1-cm-wide samples at every 1 cm for the top 63 cm and then every 4 cm down to 200 cm in 56JPC, and every 1 cm down to 71 cm in 48JPC. Samples were analysed at Cardiff University on a Beckman Coulter Multisizer 4 using the Enhanced Performance Multisizer 4 beaker and stirrer setting 30 to ensure full sediment suspension. Two or three separate aliquots were analysed for each sample, sizing 70,000 particles per aliquot. Analytical precision was approximately 1% ($\pm 0.3\,\mu$m), while full procedural error

(based on replicates of about 25% of samples, starting from newly sampled bulk sediment) was $\pm 0.8\,\mu$m.

Temperature data and constructing the T_{sub} index. Numerous studies have suggested that AMOC variability is associated with a distinct surface or subsurface (400 m) temperature fingerprint in the North Atlantic[6,11,28,41]. However, the lack of long-term observations of the precise AMOC temperature fingerprint, and models display a range of different AMOC temperature fingerprints[9,42]. Here we focus on the T_{sub} AMOC fingerprint, proposed by Zhang[11] on the basis of covariance between a modelled AMOC, the spatial pattern of the leading mode of subsurface (400 m) temperature variability, and sea-surface height changes. These model-based relationships are supported by similar relationships (spatial and temporal) observed in recent instrumental data of subsurface temperature and sea-surface height. The agreement between our DWBC$_{LSW}$ AMOC reconstruction, observed Labrador Sea density changes, and the T_{sub} AMOC fingerprint provides support for our approach and suggests that the T_{sub} AMOC fingerprint is capturing an important component of deep AMOC variability. Differences between the various proposed AMOC temperature fingerprints probably reflects their sensitivity to different aspects of the AMOC and heat transport in the North Atlantic (for example, the AMOC versus SPG circulation[28]); the temperature response to each of these components may be resolved if more comprehensive spatial networks of past North Atlantic temperature variability are generated[43].

We selected records used in the OCEAN 2 K synthesis[44] from the Northwest Atlantic slope and the subpolar Northeast Atlantic, and supplemented them with additional records that also record past temperature variability in the subsurface ocean of the chosen region. We excluded cores that did not have a modern core-top age (AD 1950 or younger) or a resolution of better than 100 years. We selected foraminiferal-based temperature proxies because they record subsurface temperatures (typically at 50–200 m depth), upon which the T_{sub} proxy is based. We avoided other temperature proxies (for example, alkenones, coccolithores and diatoms) that are typically more sensitive to sea-surface temperature, rather than to T_{sub}, and which also use the fine fraction that—at the drift sites required for the necessary age resolution—contains substantial allochthonous material, compromising the fidelity of in situ temperature reconstruction[45,46].

We normalized all T_{sub} records to the interval AD 1750–2000 (the length of the shortest records). We calculated the T_{sub} proxy reconstruction as the difference between the stacked temperature records of the Northwest and Northeast Atlantic. Our results are insensitive to the precise binning or stacking method (Extended Data Fig. 4). The sedimentation rates of the cores used, combined with the effects of bioturbation, mean we cannot resolve signals on timescales shorter than about 20–50 years. Age model uncertainty is estimated to be up to about 30 years for the past 150 years or so (where cores can be dated on the basis of ^{210}Pb signatures), and around 100 years for AD 400–1800 (where ^{14}C dating is relied upon). Therefore, the optimal bin intervals chosen were 50 years for AD 1800–2000, and 100 years for AD 400–1800. Results using just 50-year and 100-year bins, as well as 30-year bins for the top 200 years, are shown in Extended Data Fig. 4.

Data availability. The proxy data that support these findings are provided as Source Data for Figs. 2 and 3 and Extended Data Figs. 1, 2, 4, 5, 6 and 9, and at National Geophysical Data Center (NGDC) Paleoclimatology database (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets). Model data are available from J.I.R. (j.i.robson@reading.ac.uk) upon reasonable request.

34. Megann, A. et al. GO5.0: the joint NERC–Met Office NEMO global ocean model for use in coupled and forced applications. *Geosci. Model Dev* **7**, 1069–1092 (2014).
35. Williams, K. D. et al. The Met Office Global Coupled model 2.0 (GC2) configuration. *Geosci. Model Dev* **8**, 1509–1524 (2015).
36. Shaffrey, L. C. et al. U.K. HiGEM: the new U.K. high-resolution global environment model—model description and basic evaluation. *J. Clim.* **22**, 1861–1896 (2009).
37. Bakker, P., Govin, A., Thornalley, D. J. R., Roche, D. M. & Renssen, H. The evolution of deep-ocean flow speeds and δ^{13}C under large changes in the Atlantic overturning circulation: toward a more direct model-data comparison. *Paleoceanography* **30**, 95–117 (2015).
38. Toole, J. M., Andres, M., Le Bras, I. A., Joyce, T. M. & McCartney, M. S. Moored observations of the Deep Western Boundary Current in the NWAtlantic: 2004–2014. *J. Geophys. Res. Oceans* **122**, 7488–7505 (2017).
39. Rose, N. L. Spheroidal carbonaceous fly ash particles provide a globally synchronous stratigraphic marker for the Anthropocene. *Environ. Sci. Technol.* **49**, 4155–4162 (2015).
40. McCave, I. N., Manighetti, B. & Robinson, S. G. Sortable silt and fine sediment size/composition slicing: parameters for palaeocurrent speed and palaeoceanography. *Paleoceanography* **10**, 593–610 (1995).
41. Dima, M. & Lohmann, G. Evidence for two distinct modes of large-scale ocean circulation changes over the last century. *J. Clim.* **23**, 5–16 (2010).
42. Muir, L. C. & Fedorov, A. V. How the AMOC affects ocean temperatures on decadal to centennial timescales: the North Atlantic versus an interhemispheric seesaw. *Clim. Dyn.* **45**, 151–160 (2015).

43. Ortega, P., Robson, J., Moffa-Sanchez, P., Thornalley, D. J. R. & Swingedouw, D. A last millennium perspective on North Atlantic variability: exploiting synergies between models and proxy data. *CLIVAR Exch.* **72**, 61–67 (2017).
44. McGregor, H. V. et al. Robust global ocean cooling trend for the pre-industrial Common Era. *Nat. Geosci.* **8**, 671–677 (2015).
45. McCave, I. N. A poisoned chalice? *Science* **298**, 1186–1187 (2002).
46. Filippova, A., Kienast, M., Frank, M. & Schneider, R. R. Alkenone paleothermometry in the North Atlantic: a review and synthesis of surface sediment data and calibrations. *Geochem. Geophys. Geosyst.* **17**, 1370–1382 (2016).
47. Marchitto, T. & deMenocal, P. Late Holocene variability of upper North Atlantic Deep Water temperature and salinity. *Geochem. Geophys. Geosyst.* **4**, 1100 (2003).
48. Keigwin, L. D., Sachs, J. P. & Rosenthal, Y. A. 1600-year history of the Labrador Current off Nova Scotia. *Clim. Dyn.* **21**, 53–62 (2003).
49. Keigwin, L. D. & Pickart, R. S. Slope water current over the Laurentian Fan on interannual to millennial time scales. *Science* **286**, 520–523 (1999).
50. Genovesi, L. et al. Recent changes in bottom water oxygenation and temperature in the Gulf of St. Lawrence: micropaleontological and geochemical evidence. *Limnol. Oceanogr.* **56**, 1319–1329 (2011).
51. Hall, I. R., Boessenkool, K. P., Barker, S., McCave, I. N. & Elderfield, H. Surface and deep ocean coupling in the subpolar North Atlantic during the last 230 years. *Paleoceanography* **25**, PA2101 (2010).
52. Moffa-Sanchez, P., Born, A., Hall, I. R., Thornalley, D. J. R. & Barker, S. Solar forcing of North Atlantic surface temperature and salinity over the past millennium. *Nat. Geosci.* **7**, 275–278 (2014).
53. Thornalley, D. J. R., Elderfield, H. & McCave, I. N. Holocene oscillations in temperature and salinity of the surface subpolar North Atlantic. *Nature* **457**, 711–714 (2009).
54. Richter, T. O., Peeters, F. J. C. & van Weering, T. C. E. Late Holocene (0-2.4 ka BP) surface water temperature and salinity variability, Feni Drift, NE Atlantic Ocean. *Quat. Sci. Rev.* **28**, 1941–1955 (2009).
55. Morley, A. et al. Solar modulation of North Atlantic central water formation at multidecadal timescales during the late Holocene. *Earth Planet. Sci. Lett.* **308**, 161–171 (2011).
56. Morley, A., Rosenthal, Y. & deMenocal, P. Ocean-atmosphere climate shift during the mid-to-late Holocene transition. *Earth Planet. Sci. Lett.* **388**, 18–26 (2014).
57. Sicre, M.-A. et al. A 4500-year reconstruction of sea surface temperature variability at decadal time-scales off North Iceland. *Quat. Sci. Rev.* **27**, 2041–2047 (2008).
58. Joyce, T. M. & Zhang, R. On the path of the Gulf Stream and the Atlantic Meridional Overturning Circulation. *J. Clim.* **23**, 3146–3154 (2010).
59. Suman, D. O. & Bacon, M. P. Variations in Holocene sedimentation in the North American Basin determined from ^{230}Th measurements. *Deep-Sea Res.* **36**, 869–878 (1989).
60. Adkins, J. F., Boyle, E. A., Keigwin, L. & Cortijo, E. Variability of the North Atlantic thermohaline circulation during the last interglacial period. *Nature* **390**, 154–156 (1997).
61. Hodson, D. L. R., Robson, J. I. & Sutton, R. T. An anatomy of the cooling of the North Atlantic Ocean in the 1960s and 1970s. *J. Clim.* **27**, 8229–8243 (2014).
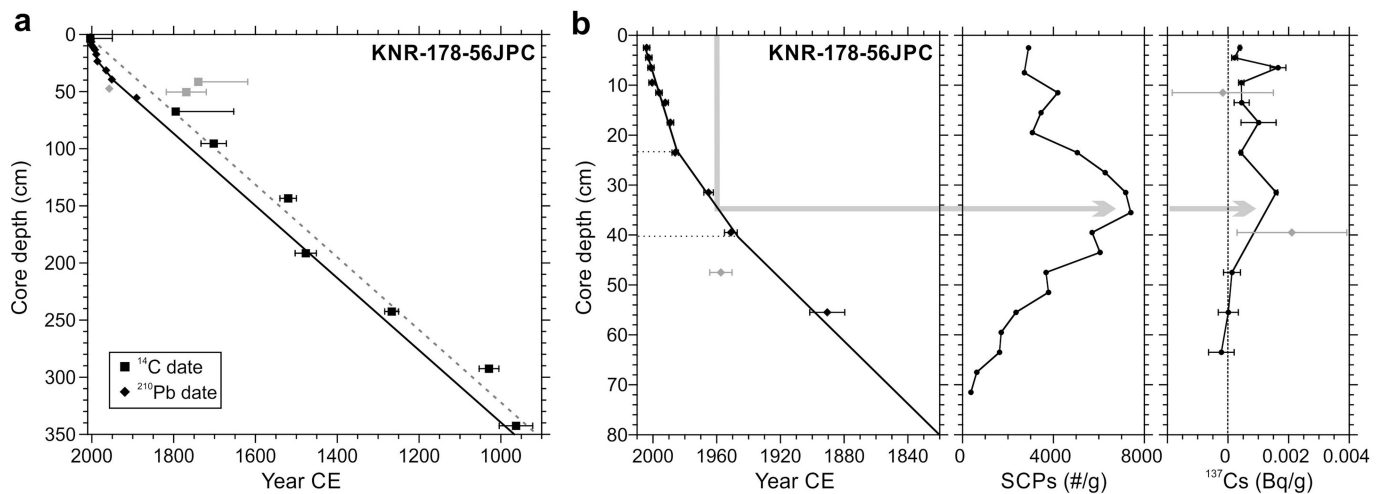
Extended Data Fig. 1 | Age model for core KNR-178-56JPC. a, ^{14}C and ^{210}Pb dating. The ^{14}C ages (with 1σ ranges; grey, rejected dates) from planktic foraminifera yield a modern core-top age and indicate an average sedimentation rate over the past 1,000 years of 320 cm kyr[−1] (dashed line). The presence throughout the core of abundant lithogenic grains in the >150-µm fraction—along with the coarse sortable-silt mean grain size values—suggests that some reworking of foraminifera has probably occurred, resulting in average ^{14}C ages that may be slightly (around 50 years) older than their final depositional age, consistent with the fact that the ^{210}Pb dates do not splice smoothly into the ^{14}C ages (the ^{14}C ages appear slightly too old). The final age model was therefore based on the ^{210}Pb ages for the past century, and was then simply extrapolated back in time using the linear sedimentation rate of 320 cm kyr[−1]. Given that none of our findings depend on close age control in the older section of this core (that is, before AD 1880), this uncertainty (with converted ^{14}C ages being about 50 years older than the extrapolated linear age model) does not affect our conclusions. **b**, Left, the age model for the top 80 cm of core 56JPC is based on ^{210}Pb dating of bulk sediment, using the constant initial concentration (CIC) method (rejecting the date at 47 cm, which probably indicates a burrow). A simple two-segment linear fit to the ^{210}Pb dates is adopted (rather than point-to-point interpolation or a spline) because sedimentological evidence—an abrupt increase in the percentage of coarse fraction at 23 cm depth, not observed elsewhere in the core—is indicative of a step change in the sedimentation rate. Horizontal dashed lines denote the depths of the segments at which the sedimentation rate is inferred to change. Centre, further support for the age model of 56JPC over the past century comes from the down-core abundance profile of spheroidal carbonaceous particles (SCPs, derived from high-temperature fossil fuel combustion, counted as described[39]), which ramped up from the mid to late 1800s and peaked in the 1950s to 1970s (40 cm to 25 cm) before declining over recent decades, consistent with the ^{210}Pb-based age model. Right, the occurrence of ^{137}Cs in the top 40 cm or so of the core is also consistent with the ^{210}Pb-based age of around 1950 at 40 cm. The age uncertainty (1σ) for the past 60 years of the core is estimated at ±2–3 years. We note that the sediment core top is at 3 cm depth in the core-liner.
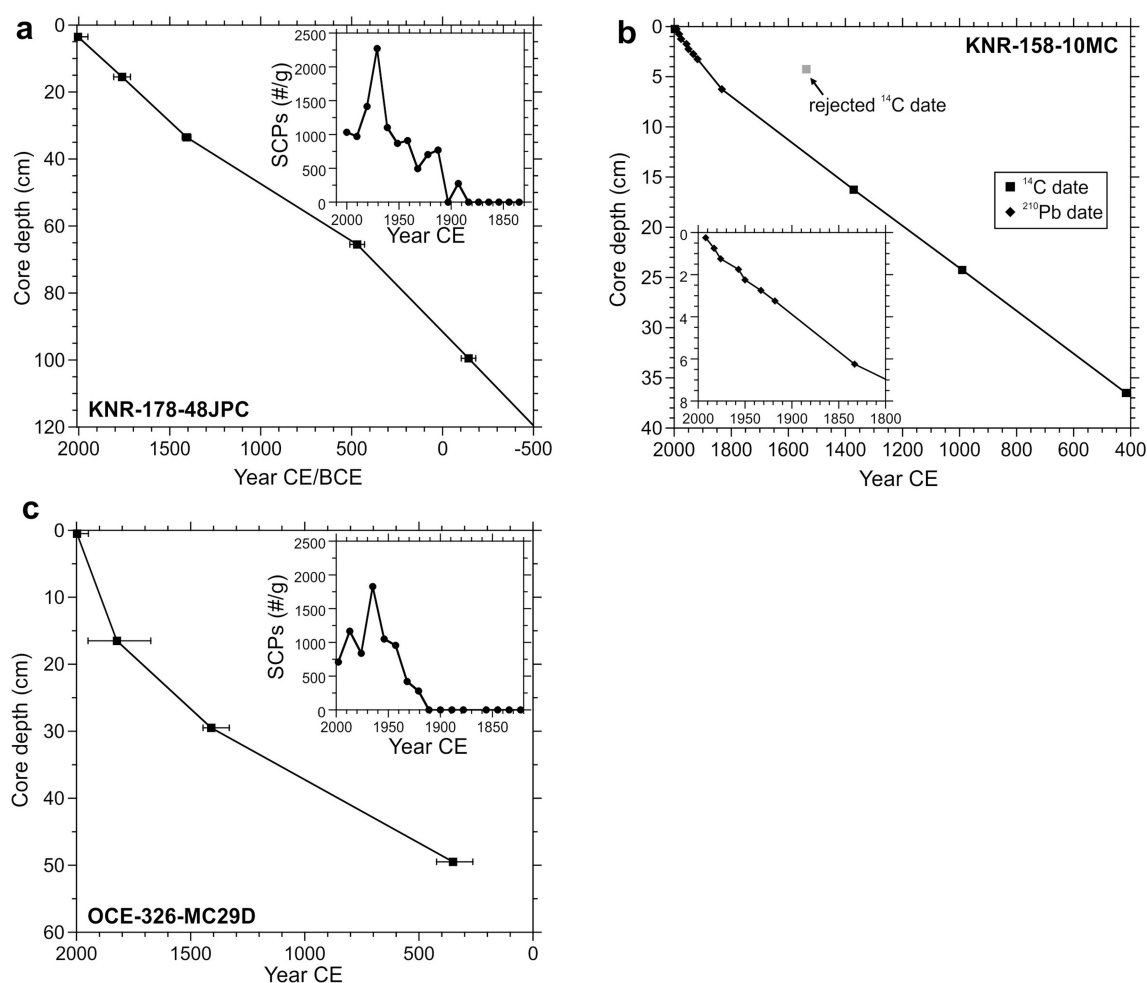
Extended Data Fig. 2 | Age models for additional cores. a, ^{14}C-based age model, derived from linear interpolation of ^{14}C-dated planktic foraminifera (with 1σ ranges) in sediment 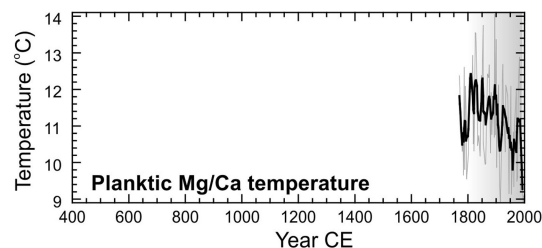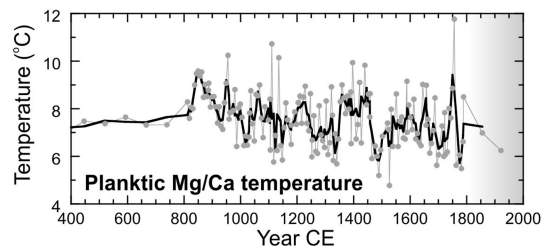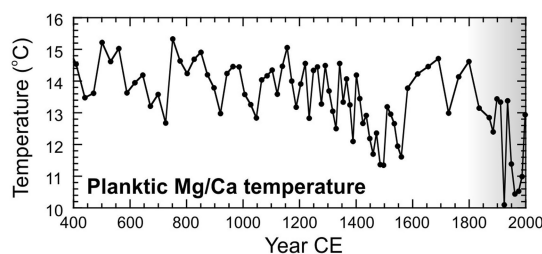core KNR-178-48JPC (used for the DWBC$_{LSW}$ sortable-silt reconstruction), yielding a modern core-top age and an average sedimentation rate of around 50 cm kyr^{-1}. We note that the core top is at 3 cm depth in the core-liner. The inset shows the SCP profile for 48JPC on the basis of the ^{14}C age model, confirming the modern age of the top sediments, with SCPs showing the expected profile— increasing in concentration from the late 1800s onwards, peaking at around 1950 to 1970, and declining afterwards. **b**, Updated age model for core KNR-158-10MC (after ref. [47]; used in Extended Data Fig. 5 examining regional near-surface temperature trends in the Northwest Atlantic during the industrial era), using new ^{210}Pb dating (CIC method) for the top 7 cm and rejecting the anomalously old ^{14}C age at 4 cm depth; the inset shows

[210]Pb age constraints in the top 8 cm. A single detectable occurrence of ^{137}Cs at 2–2.5 cm (equivalent to 1957 on the ^{210}Pb-based age model) can be linked to the bomb peak at 1963, supporting the age model. Also, SCPs were found in the top 5 cm of this core, confirming the industrial-era age for the top 5 cm; however, the low concentrations of SCPs prevent meaningful interpretation of the down-core trends and are not shown. **c**, Age model for core OCE-326-MC29B (used for T_{sub} reconstruction of the Northwest Atlantic shelf): [14]C ages of planktic foraminifera (with 1σ ranges), from ref. [48]. Support for this age model is provided by the SCP concentrations (inset; this study), which show the expected down-core profile[39] when plotted using the ^{14}C ages. [210]Pb dating[48] also suggests a sedimentation rate of around 120 cm kyr^{-1} for the uppermost sediments, consistent with the ^{14}C ages and SCP profile.

NW ATLANTIC SLOPE

a Emerald Basin (29MC, 250m water depth, site 2)

Planktic δ¹⁸O

% Polar species

Benthic δ¹⁸O

Benthic Mg/Ca temperature
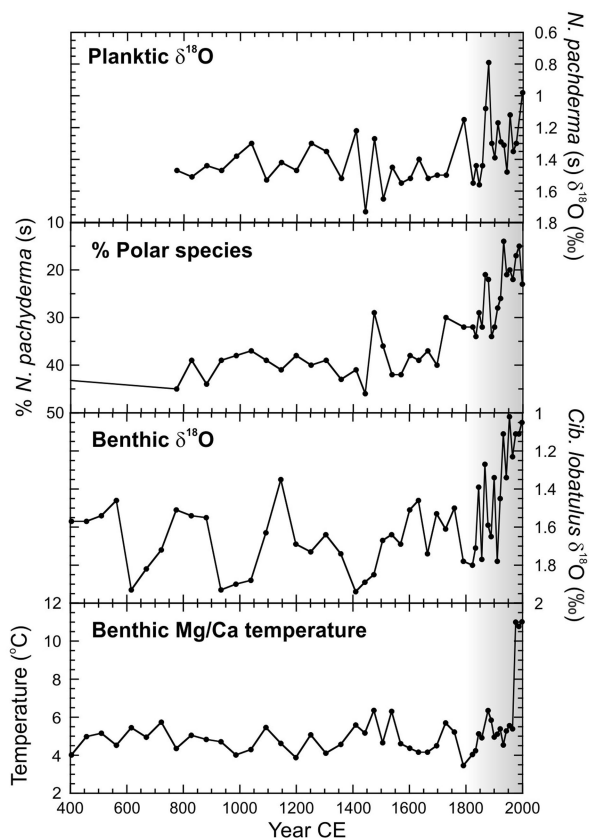
b Laurentian Fan (13MC, site 3)

% Polar species
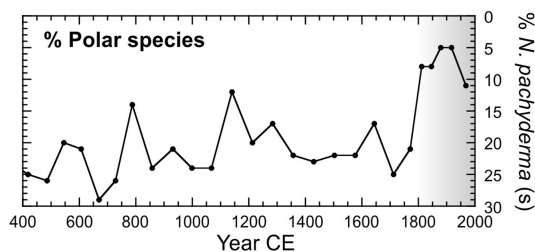
c Gulf of St Lawrence (409m water depth, site 1)

Benthic δ¹⁸O
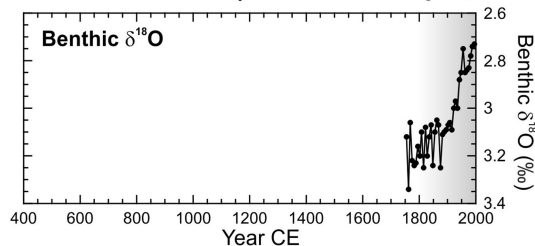
NE ATLANTIC SPG

d Gardar drift (site 4)

Planktic Mg/Ca temperature

e Bjorn drift (site 5)

Planktic Mg/Ca temperature

f Feni drift (site 6)

Planktic Mg/Ca temperature

g ENACW, formed in eastern SPG (899m water depth, site 7)

Benthic Mg/Ca temperature
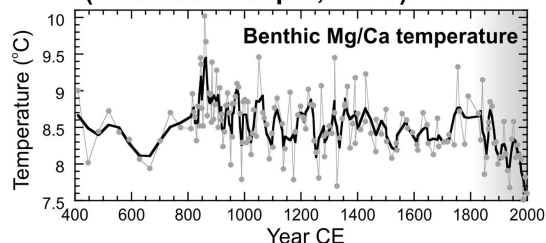
h *North Iceland shelf/Rahmstorf SPG SST

Alkenone SST
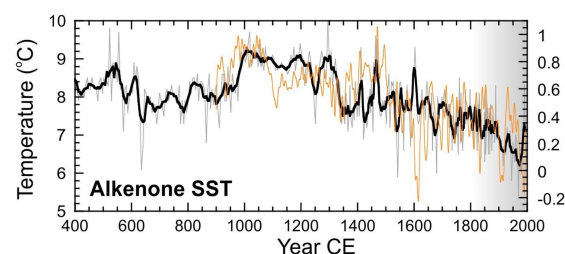
Extended Data Fig. 3 | Raw data for construction of the T_{sub} AMOC proxy shown in Fig. 3. Locations are shown in Fig. 2b. **a–c,** Temperature proxy records[48–50] used for the Northwest Atlantic stack (Emerald Basin, Laurentian Fan and Gulf of St Lawrence), where model studies[11,12] indicate that AMOC weakening results in warming of surface and subsurface waters. **d–g,** Records used to reconstruct Northeast Atlantic SPG subsurface temperatures: **d,** Gardar drift[51]; **e,** combined South Iceland data (Bjorn drift)[52,53]; **f,** Feni drift[54]; **g,** Eastern North Atlantic Central

Water (ENACW), largely composed of waters formed in the eastern SPG[55,56]. **h,** The high-resolution alkenone sea-surface temperature (SST) record from the North Iceland shelf[57] was not included because it is not located within the open North Atlantic SPG (although it does also show, like the other Northeast Atlantic records, that the lowest temperature of the past 1,600 years occurred during the most recent century). Also shown for reference is the Rahmstorf central SPG SST reconstruction (based largely on terrestrial proxies)[6].
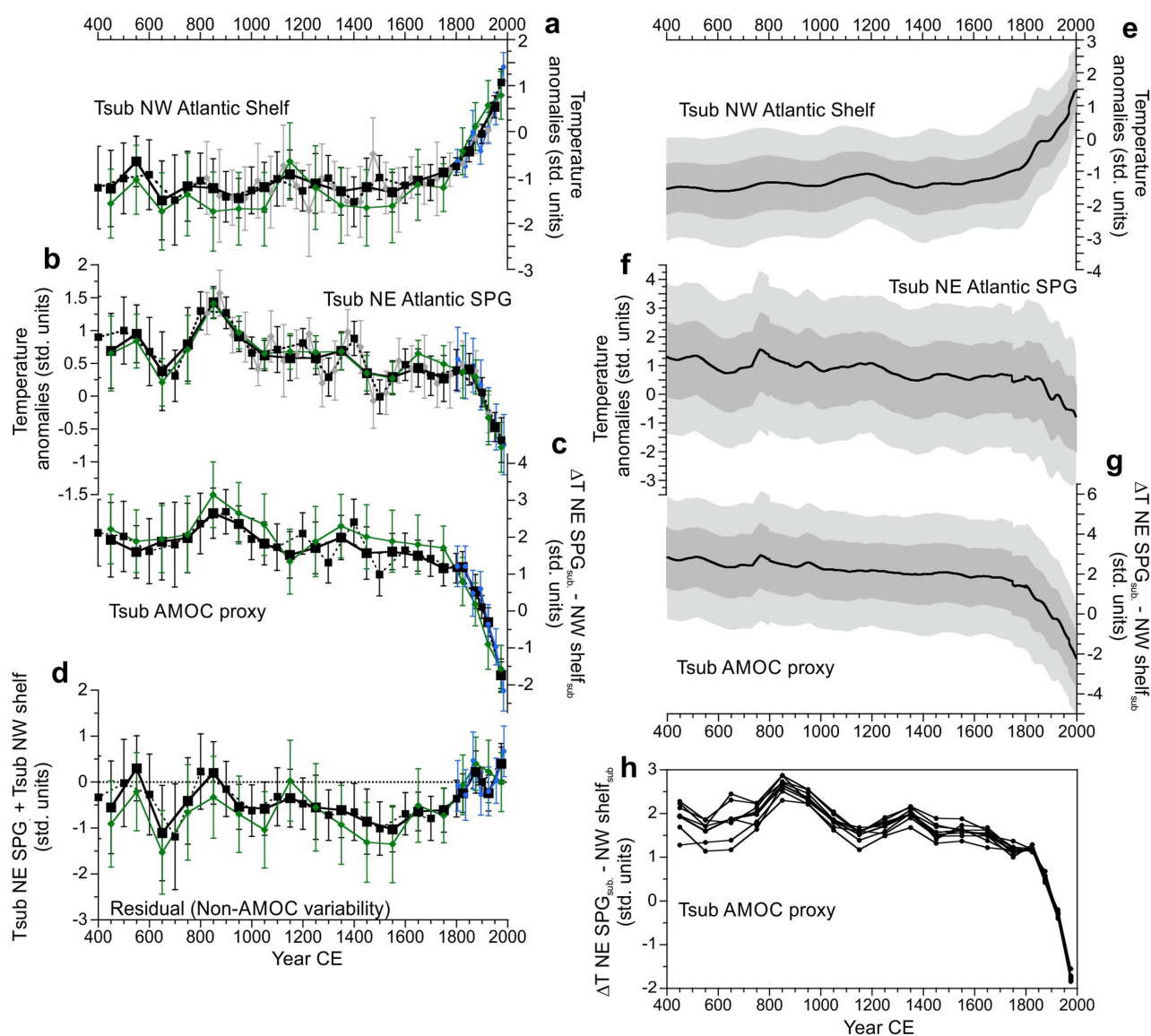
Extended Data Fig. 4 | Different binning and averaging approaches and the residual temperature signal. a, **b**, Stacked, normalized proxy temperature data (T_{sub}) from the Northwest Atlantic shelf/slope (**a**) and Northeast Atlantic SPG (**b**). **c**, The derived T_{sub} AMOC proxy, calculated as the numerical difference between the stacks shown in **a** and **b**. **d**, The residual temperature variability in stacks **a** and **b** that is not described by the (anti-phased dipole) T_{sub} AMOC proxy shown in **c**—that is, the in-phase temperature variability common to both stacks, calculated as the numerical sum of the two stacks (if divided by two, this would be the numerical mean). This represents the inferred non-AMOC-related temperature variability common to both regions, and broadly resembles Northern Hemisphere temperature reconstructions, most notably colder residual temperatures during the LIA, around 1350 to 1850. For **a–d**, black solid lines and squares represent preferred binning (50 years for 1800–2000;

100 years for 400–1800); green line and symbols, as for preferred binning, but with stacks produced by first binning the proxy data at each site and then averaging these binned site values, as opposed to binning all the proxy data together in one step (the former ensures equal weighting for each site, the latter biases the final result to the higher-resolution records); black dashed lines and symbols, 100-year bins offset by 50 years from the preferred bins; grey lines and symbols, 50-year bins (not shown for **c** and **d**); blue lines and symbols; 30-year bins for 1790–2000. Error bars for **a–d** are ±2 s.e. **e–g**, As for **a–c**, except using a Monte Carlo approach and published uncertainties for age assignment and temperature reconstructions; light and dark grey shading represent ±1σ and ±2σ, respectively. **h**, Jackknife version of **c**, with each line representing the T_{sub} AMOC proxy but leaving out one of the individual proxy records each time.

a

(AMOC -ve) - (AMOC +ve), Threshold +/- 3 Sv
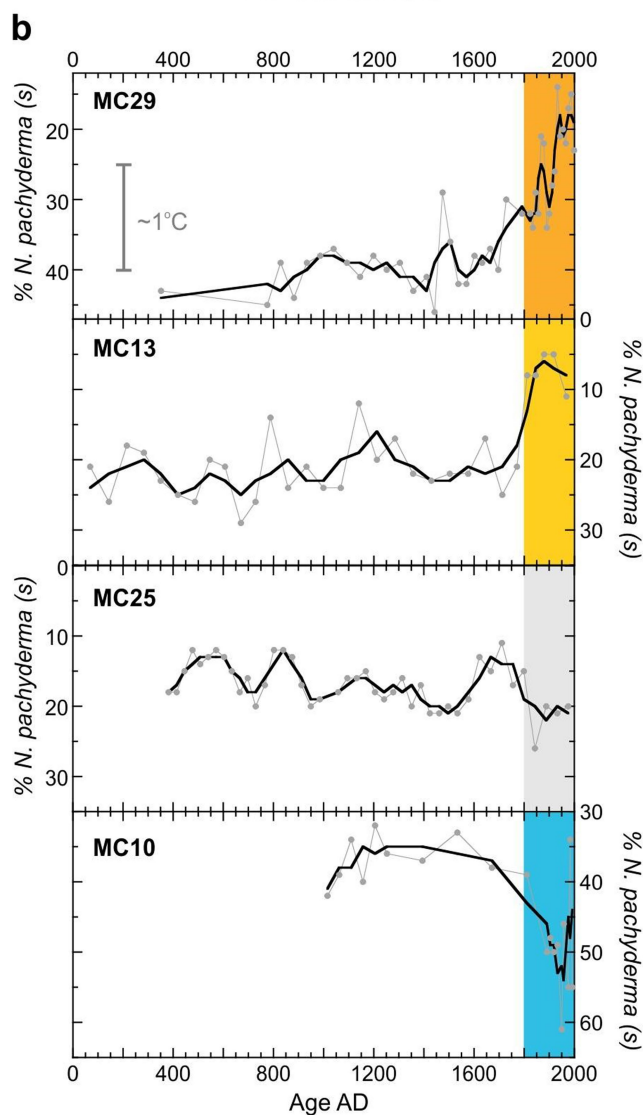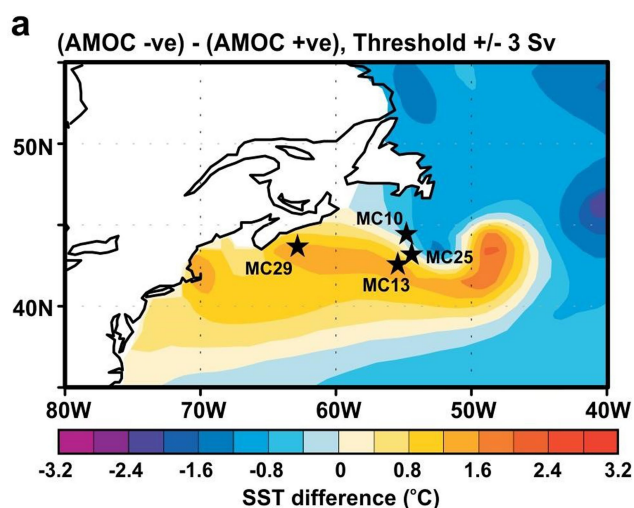
SST difference (°C)

Extended Data Fig. 5 | SST response of the Northwest Atlantic to AMOC weakening. a, Modelled SST difference between a weak (negative) and strong (positive) AMOC[58]. This pattern is model-dependent, with the study cited here[58] chosen because of its good agreement with observations of Gulf Stream variability. The locations of cores used for panel **b** are shown by black stars. **b**, Percentage abundances of the polar species *N. pachyderma* (sinistral) in marine sediment cores from the Northwest Atlantic, as an indicator of near-surface (around 75 m) temperatures. A 15% increase indicates around 1 °C of cooling (we note the reversed *y* axes). The opposing trends over the past 200 years are consistent with the SST pattern modelled for a weakening of the AMOC, as shown in panel **a**. Data and age models for the cores are: OCE326-MC29[48] using the original ^{14}C dating and as shown in Extended Data Fig. 2; OCE326-MC13 and OCE326-MC25[49] using the original ^{14}C age ties at the top and bottom of the core and scaling the intervening sedimentation rate to the percentage of $CaCO_3$ content[49,59,60]; KNR158-MC10, this study, using the age model in Extended Data Fig. 2.

b
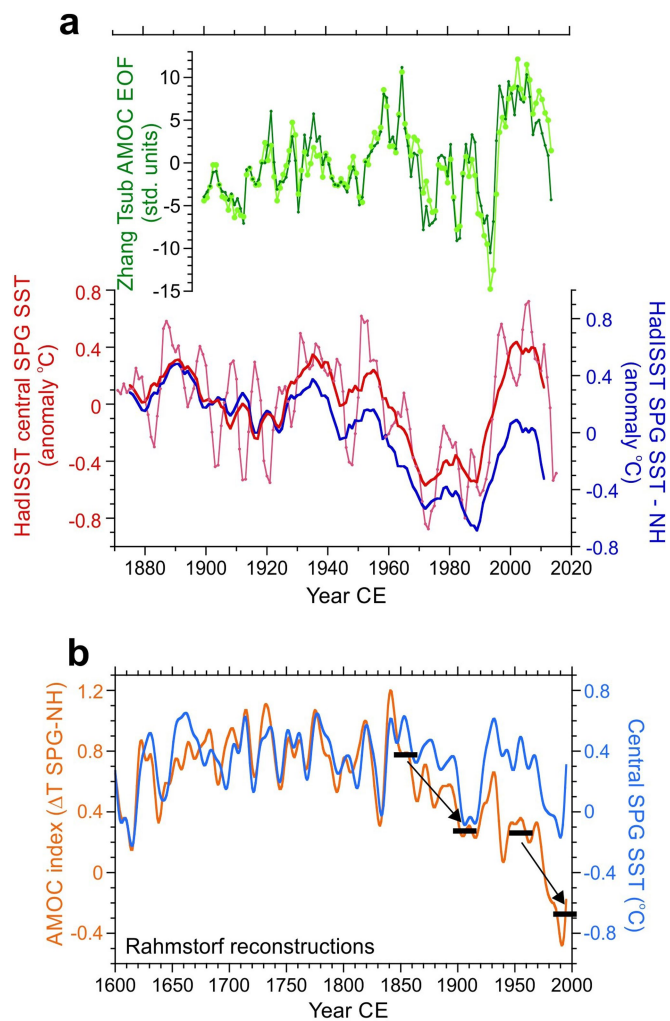
Extended Data Fig. 6 | Temperature fingerprints of the AMOC during the twentieth century. a, Top, T_{sub} AMOC fingerprint[11] obtained using empirical orthogonal function (EOF) analysis of the EN4 dataset (light green, the leading mode (EOF1) of T_{sub} variability from 1993–2003, as defined by Zhang[11], applied to the EN4 data; dark green, the second mode of T_{sub} variability (EOF2) of the North Atlantic for 1900–2015, equivalent to the EOF1 defined for 1993–2003). No substantial twentieth-century AMOC decline is seen in this observation-based reconstruction. Bottom, instrument-based reanalysis of the 'cold blob' central SPG region (red; 3-year (thin line) and 11-year (thick line) smoothing; 47° N to 57° N, 30° W to 45° W) used in the Rahmstorf SST AMOC proxy[6]. The data are from the HadISST project. The reconstructed central SPG SST bears some resemblance to the T_{sub} AMOC fingerprint record, which is not unexpected given that the central SPG forms a substantial spatial component of the T_{sub} fingerprint. No clear decrease is shown in the central SPG SST, and the equivalent Rahmstorf AMOC proxy[6] (blue; central SPG minus the Northern Hemisphere (NH) temperature) declines during the twentieth century because of the subtraction of the NH warming trend. **b**, Reconstructed (predominantly terrestrial-based) AMOC proxy (orange; the temperature difference between the central SPG and the NH) and the central SPG SST reconstruction[6] (blue). There is a two-step decline in the AMOC proxy, at 1850–1900 and 1950–2000—the former being mainly the result of a strong cooling of the SPG (which probably weakened northward heat transport, paralleling the weakening shown by our DWBC proxy), and the latter being due mainly to subtraction of the strong NH warming trend, rather than a persistent SPG cooling.

GC2 climatological Surface Currents
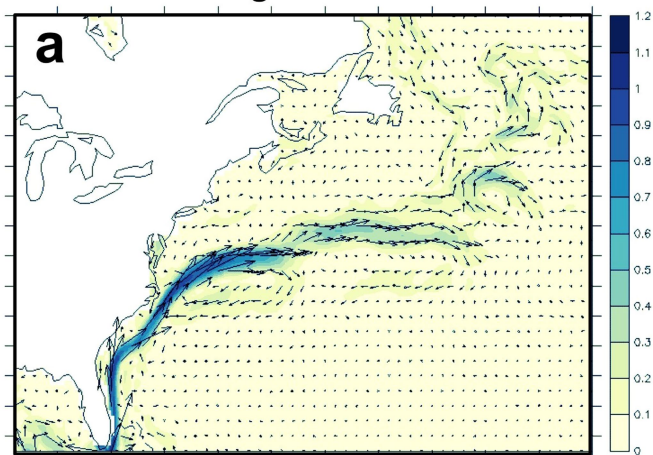
Observed (OSCAR) Surface Currents
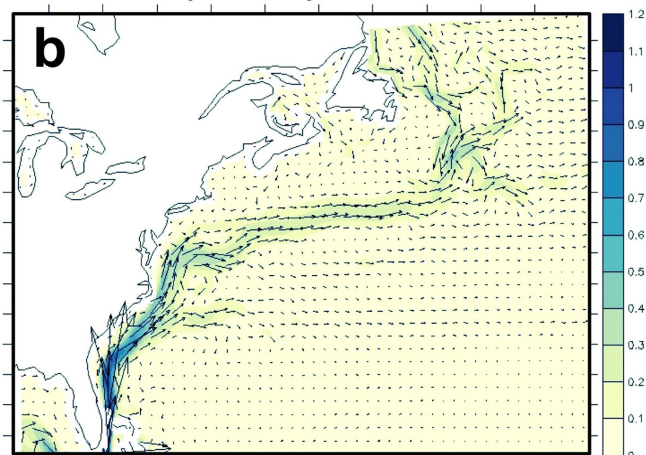
Extended Data Fig. 7 | DWBC changes in model HadGEM3-GC2. a, b, Climatological surface current direction (in arrows) and speed (shaded, m s^{-1}) obtained from the control simulation with HadGEM3-GC2 and the satellite product OSCAR.
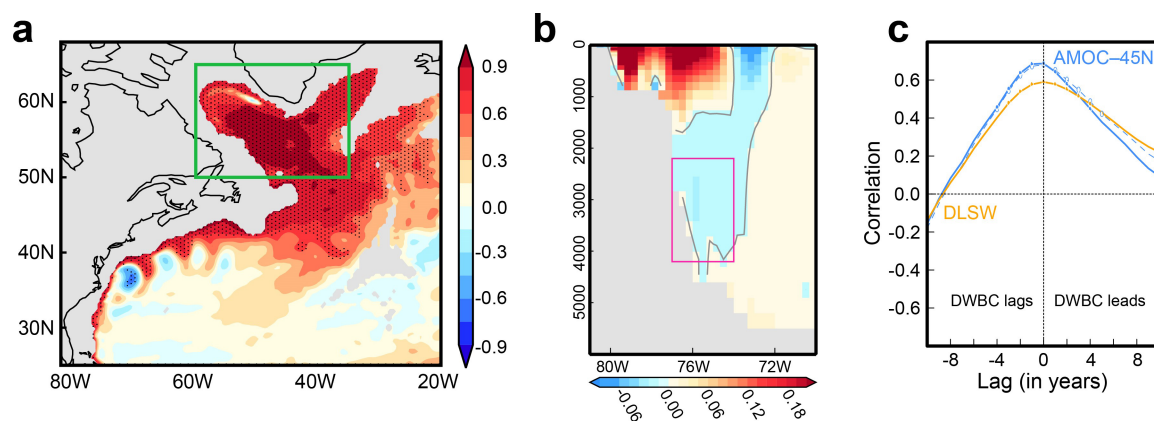
Extended Data Fig. 8 | Modelled link between DWBC velocity, DLSD and AMOC in the HiGEM model. a, Correlation (colour bar) of the vertically averaged ocean density (at 1,000–2,500 m) with the DLSD index (as defined in ref. [4]; green box, 1,000–2,500 m average) in a 340-year present-day control run of the HiGEM model (see ref. [36]). **b**, Climatology of the modelled meridional ocean velocity (in m s^{-1}) averaged between 30° N and 35° N, illustrating the modelled position of the DWBC. The

y axis shows the water depth in metres. **c**, Cross-correlations between the modelled average DWBC flow speed in the pink box in panel **b** and indices of DLSD and AMOC at 45° N (the dashed line omits the Ekman component). We note that the box over which the DWBC flow index in panel **c** is averaged has changed with respect to Fig. 1, in order to take into account of the fact that the return flow is deeper in the HiGEM model than in HadGEM3-GC2.
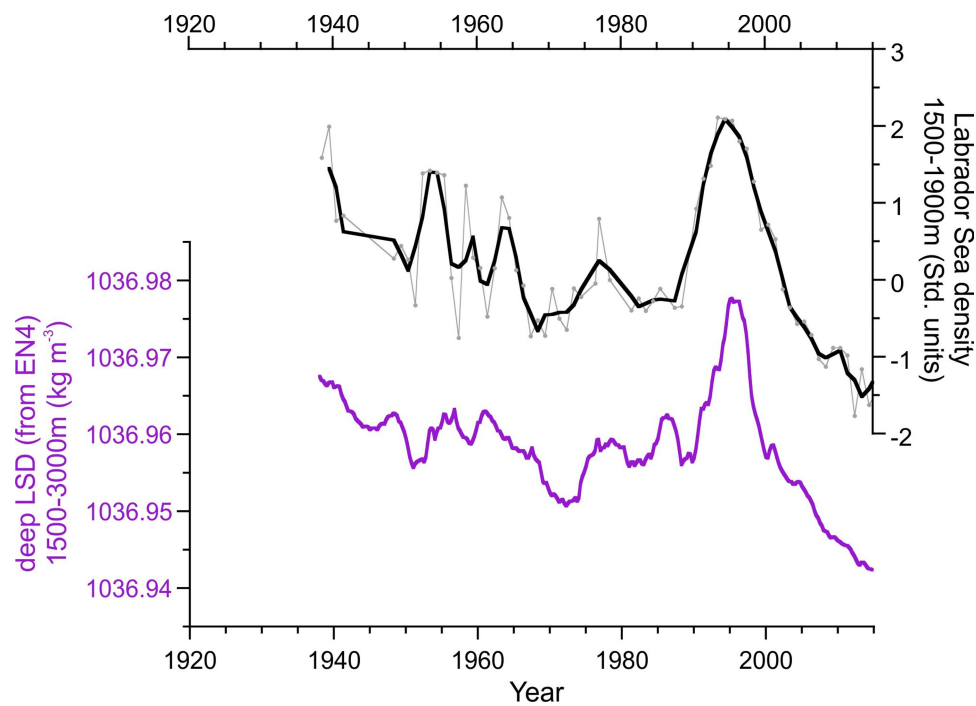
Extended Data Fig. 9 | Comparison of Labrador Sea density parameters. The model-based DLSD parameter—proposed in ref. [4] and using the EN4 reanalysis dataset—incorporates a larger area and greater depth range than do instrumental-data-only studies, such as ref. [5], which examines past variability in Labrador Sea convection and focuses on the central Labrador Sea and on depths less than 2,000 m, where most observational data are available. The comparison here of DLSD (purple line, three-year mean) from the EN4 dataset with instrumental data on density changes in the central Labrador Sea at 1,500–1,900 m depth (grey line, annual averages; black line, three-year mean) illustrates that the two parameters show very similar variability. Both are dominated by the density changes caused by deep convection in the Labrador Sea, which can reach down to around 2,000 m. Estimates of uncertainty are discussed in ref. [61].
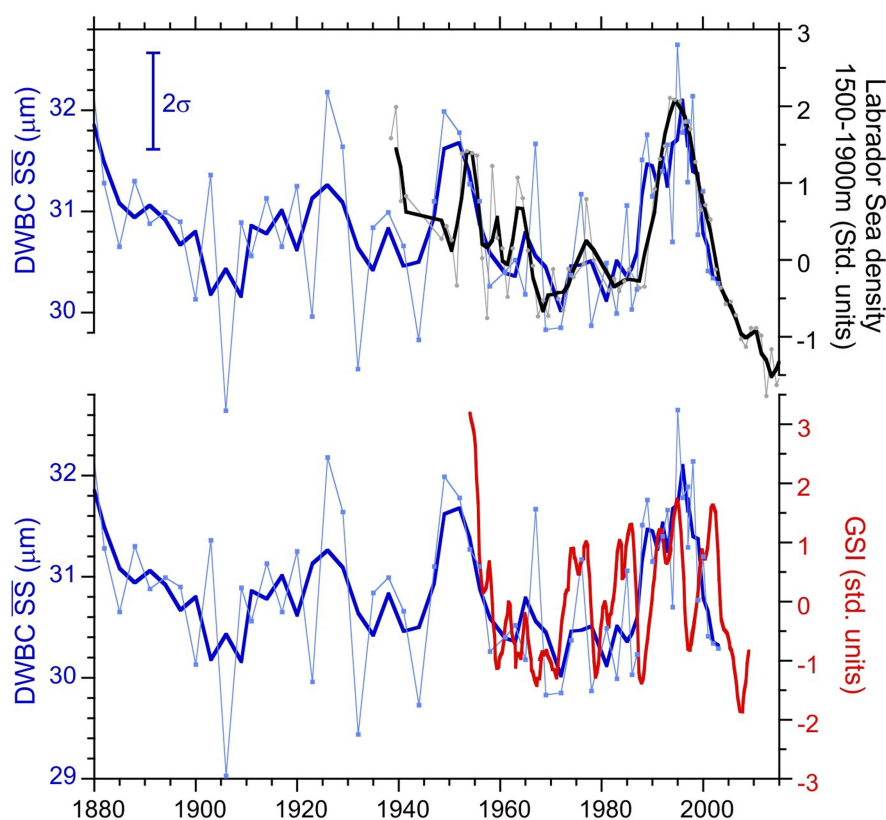
Extended Data Fig. 10 | Comparison with Gulf Stream Index (GSI).
A direct influence of the changing position of the Gulf Stream on the grain size of our core sites can be ruled out by comparing instrumental records of the Gulf Stream position (red, GSI[58]) with the down-core sortable-silt (SS) mean grain size data in 56JPC (blue; thicker line is three-point smoothed). There is no clear correlation between these two proxies (bottom). However, there is a coupling between our SS data (which represent inferred $DWBC_{LSW}$ flow speed) and density changes in the deep Labrador Sea (grey, annual; black, three-point smoothed; top panel). The 2σ SS error bar ($n = 30$) is for the three-point mean.

LETTER

Accelerated increase in plant species richness on mountain summits is linked to warming

Manuel J. Steinbauer[1,2]*, John-Arvid Grytnes[3], Gerald Jurasinski[4], Aino Kulonen[3,5], Jonathan Lenoir[6], Harald Pauli[7,8], Christian Rixen[5], Manuela Winkler[7,8], Manfred Bardy-Durchhalter[7,8], Elena Barni[9], Anne D. Bjorkman[1,10,11], Frank T. Breiner[12,13], Sarah Burg[5], Patryk Czortek[14], Melissa A. Dawes[5,13], Anna Delimat[15], Stefan Dullinger[16], Brigitta Erschbamer[17], Vivian A. Felde[3], Olatz Fernández-Arberas[18], Kjetil F. Fossheim[3], Daniel Gómez-García[18], Damien Georges[1,19], Erlend T. Grindrud[20], Sylvia Haider[11,21], Siri V. Haugum[3], Hanne Henriksen[20], María J. Herreros[18], Bogdan Jaroszewicz[14], Francesca Jaroszynska[3,22], Robert Kanka[23], Jutta Kapfer[24], Kari Klanderud[20], Ingolf Kühn[11,21,25], Andrea Lamprecht[7,8], Magali Matteodo[5,26], Umberto Morra di Cella[27], Signe Normand[1,28], Arvid Odland[29], Siri L. Olsen[30], Sara Palacio[18], Martina Petey[27], Veronika Piscová[23], Blazena Sedlakova[31], Klaus Steinbauer[7,8], Veronika Stöckli[5,32], Jens-Christian Svenning[1,28], Guido Teppa[9], Jean-Paul Theurillat[33,34], Pascal Vittoz[26], Sarah J. Woodin[22], Niklaus E. Zimmermann[13,35] & Sonja Wipf[5]*

Globally accelerating trends in societal development and human environmental impacts since the mid-twentieth century[1–7] are known as the Great Acceleration and have been discussed as a key indicator of the onset of the Anthropocene epoch[6]. While reports on ecological responses (for example, changes in species range or local extinctions) to the Great Acceleration are multiplying[8,9], it is unknown whether such biotic responses are undergoing a similar acceleration over time. This knowledge gap stems from the limited availability of time series data on biodiversity changes across large temporal and geographical extents. Here we use a dataset of repeated plant surveys from 302 mountain summits across Europe, spanning 145 years of observation, to assess the temporal trajectory of mountain biodiversity changes as a globally coherent imprint of the Anthropocene. We find a continent-wide acceleration in the rate of increase in plant species richness, with five times as much species enrichment between 2007 and 2016 as fifty years ago, between 1957 and 1966. This acceleration is strikingly synchronized with accelerated global warming and is not linked to alternative global change drivers. The accelerating increases in species richness on mountain summits across this broad spatial extent demonstrate that acceleration in climate-induced biotic change is occurring even in remote places on Earth, with potentially far-ranging consequences not only for biodiversity, but also for ecosystem functioning and services.

Mountains are particularly sensitive to ecological change and are experiencing some of the highest rates of warming under anthropogenic climate change[10,11]. Numerous reports of species redistribution towards summits[8,12–14] and warming-induced changes in biodiversity on summits[13,15,16] suggest that mountain biota are highly sensitive to increasing temperatures[17]. The current accelerating trend in temperature increase[1,6] should therefore also affect the velocity of

changes observed for mountain biota. Appropriate empirical assessments of the rate of change in the velocity of ecological responses (biodiversity and ecosystem trajectories) to accelerated global warming require long-term resurveys (for example, time series) of species communities, but these are scarce and localized. Mountain summits are especially suited for long-term studies of biotic responses to environmental changes because they represent natural permanent study sites that are easy to re-locate over time[18,19], thus making it possible to record reliable time series. By repeatedly resurveying alpine plant communities on 302 European mountain summits dating back as far as 1871, we generated time series for century-scale and continent-wide biodiversity dynamics to assess potential acceleration trends in plant diversity dynamics (Fig. 1). Using these time series data, we tested whether the recent acceleration of climate change is driving a similarly accelerating change in species richness on mountain summits across the continent.

We found that plant species richness has increased strongly over the past 145 years on the vast majority (87%) of Europe's summits (generalized linear mixed effects model, $P < 0.001$; Fig. 2, Extended Data Fig. 1, Extended Data Table 1) and that the increase has accelerated in the most recent years. This trend is consistent across all nine covered geographical regions, with no single region showing the opposite pattern. Across all summits, the increase in plant species richness has accelerated over time (linear mixed effects models, $P < 0.001$; Fig. 3, Extended Data Table 2), and the acceleration has been particularly pronounced during the past 20–30 years (Figs. 2, 3). Fifty years ago (1957 to 1966) the rate of increase in species number averaged 1.1 species per decade (Fig. 3), whereas during the past decade (2007 to 2016) the summits gained 5.4 additional species on average (Fig. 3). There is a positive relationship between the magnitude of increase in plant species richness and

[1]Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Aarhus, Denmark. [2]GeoZentrum Nordbayern, Department of Geography and Geosciences, Friedrich-Alexander University of Erlangen-Nürnberg (FAU), Erlangen, Germany. [3]Department of Biological Sciences, University of Bergen, Bergen, Norway. [4]Landscape Ecology, University of Rostock, Rostock, Germany. [5]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland. [6]CNRS, UMR 7058 EDYSAN, Université de Picardie Jules Verne, Amiens, France. [7]GLORIA Coordination, Institute for Interdisciplinary Mountain Research at the Austrian Academy of Sciences (ÖAW-IGF), Vienna, Austria. [8]GLORIA Coordination, Center for Global Change and Sustainability at the University of Natural Resources and Life Sciences (BOKU-gW/N), Vienna, Austria. [9]Department of Life Sciences and Systems Biology, University of Torino, Torino, Italy. [10]School of GeoSciences, University of Edinburgh, Edinburgh, UK. [11]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. [12]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. [13]Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland. [14]Białowieża Geobotanical Station, Faculty of Biology, University of Warsaw, Białowieża, Poland. [15]W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków, Poland. [16]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria. [17]Department of Botany, University of Innsbruck, Innsbruck, Austria. [18]Instituto Pirenaico de Ecología (IPE-CSIC), Huesca, Spain. [19]International Agency for Research on Cancer, Lyon, France. [20]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway. [21]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany. [22]School of Biological Sciences, University of Aberdeen, Aberdeen, UK. [23]Institute of Landscape Ecology, Slovak Academy of Sciences, Bratislava, Slovakia. [24]Department of Landscape Monitoring, Norwegian Institute of Bioeconomy Research, Tromsø, Norway. [25]Department for Community Ecology, Helmholtz Centre for Environmental Research – UFZ, Halle, Germany. [26]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland. [27]Environmental Protection Agency of Aosta Valley, Saint-Christophe, Italy. [28]Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Bioscience, Aarhus University, Aarhus, Denmark. [29]Department of Natural Sciences and Environmental Health, University College of Southeast Norway, Bø, Norway. [30]Norwegian Institute for Nature Research, Oslo, Norway. [31]Administration of the Tatra National Park, Svit, Slovakia. [32]Bergwelten 21 AG, Davos Platz, Switzerland. [33]Centre Alpin de Phytogéographie, Fondation J.-M. Aubert, Champex-Lac, Switzerland. [34]Section of Biology, University of Geneva, Chambésy, Switzerland. [35]Department of Environmental Systems Science, Swiss Federal Institute of Technology ETH, Zurich, Switzerland. *e-mail: manuel.steinbauer@fau.de; sonja.wipf@slf.ch
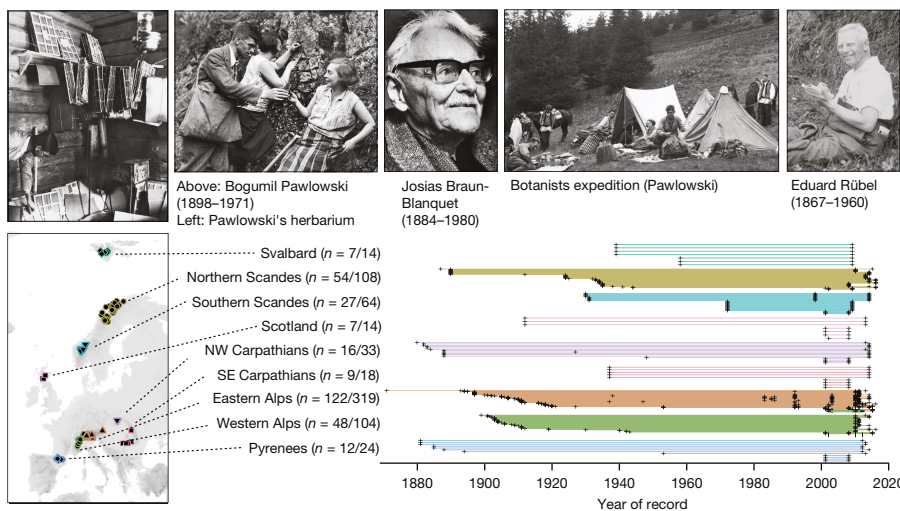
Fig. 1 | Geographical and temporal distribution of studied summits and surveys. The study is based on 698 surveys dating back to 1871 from 302 summits in nine mountain regions across Europe. Each sampled summit is indicated by one line (bottom right), with black crosses indicating survey dates. Many of the historical surveys were conducted by leading pioneers in vegetation ecology in Europe (for example, J. Braun-Blanquet, E. Du Rietz, E. Rübel and B. Pawlowski). Numbers in brackets beside the region names indicate the number of summits/surveys. Photographs reproduced with permission from ref. [31] (left, second left and second right; Botanic Garden Museum, Jagiellonian University, Kraków) and ref. [32] (middle photograph; Wiley). Right-hand figure reproduced from ref. [33]. Geospatial data for the map in all figures are from the WorldClim project (http://www.worldclim.org/).

the rate of warming across all 302 time series (linear mixed effects models, $P < 0.001$; Fig. 4a, Extended Data Fig. 2, Extended Data Table 3).

The temporal and spatial congruence between the velocity of climate change and the species accumulation rates on mountain summits across Europe corroborates the hypothesis that warming is the primary driver
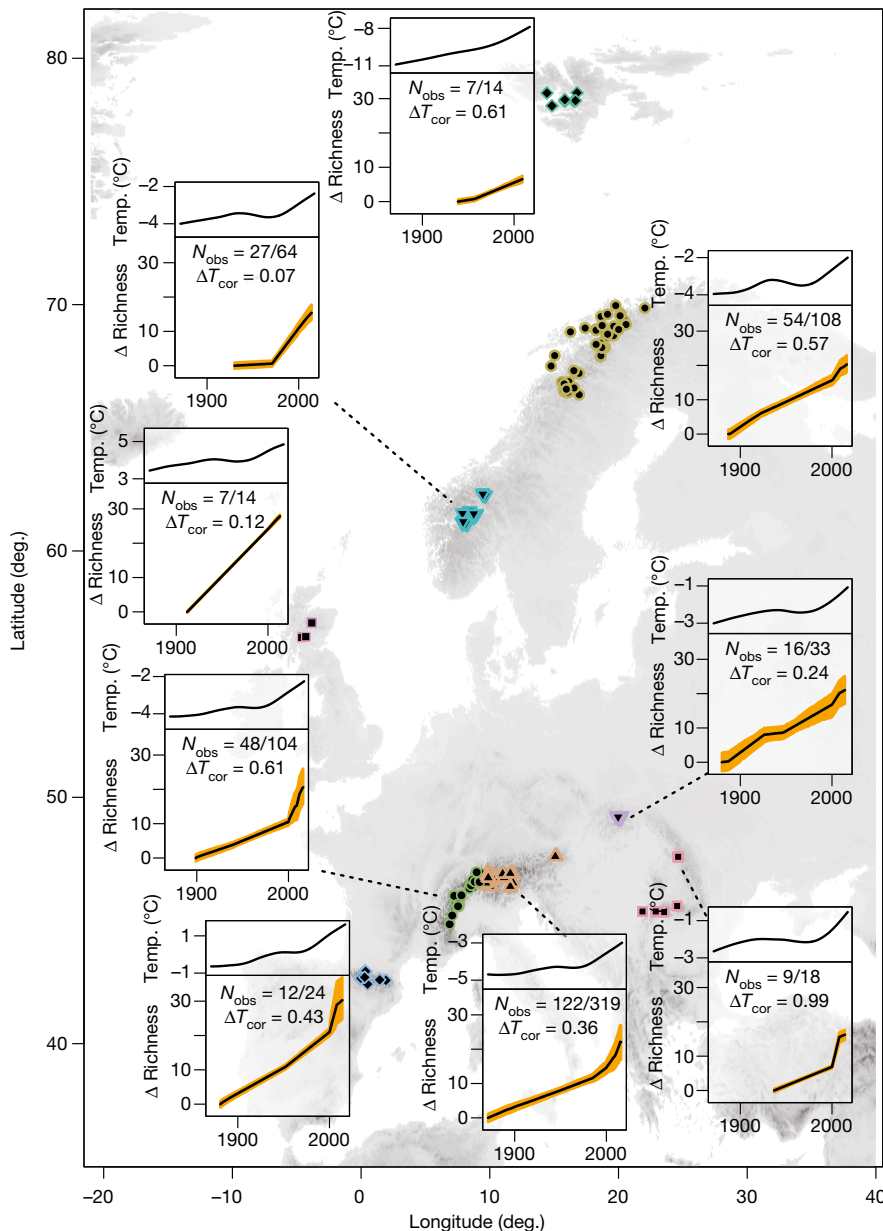
Fig. 2 | Average species richness change on mountain summits over time compared to mean annual temperature over time. Upper parts of inset panels, mean annual temperature; lower part, change in species richness (in species numbers). N_{obs}, number of summits/surveys within the mountain region providing data for the panel. Correlation between rate of change in species richness and rate of change in temperature (ΔT_{cor}) is positive for all mountain regions (Extended Data Fig. 2). Orange shading marks the 5th and 95th percentiles of the resulting richness change values from a bootstrapping approach across all summits in one region; see Extended Data Fig. 1 for methodological details.
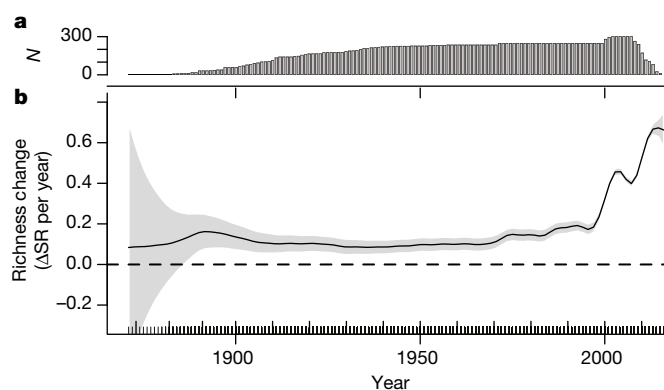
Fig. 3 | Rate of species richness change over time. a, Number of slope parameters per year (N; comparisons of earlier survey and later sampled resurvey). **b**, Rate of change in species richness (mean, black line). Positive values indicate an increase in species richness on summits and negative values indicate a decrease. Rates (ΔSR per year $= (\text{SR}_{t2} - \text{SR}_{t1})/(t_2 - t_1)$ where SR is species richness and t is time) were averaged across all summits and inversely weighted by the number of years between observations ($t_2 - t_1$) to account for temporal resolution, as a longer period between surveys might mask short-term fluctuations. The black line interpolates across all summits with a generalized additive (spline) smooth model (R package mgcv version 1.8-17; the smooth term ($k = 50$) was chosen to allow enough degrees of freedom to closely represent the underlying pattern). The shaded grey area represents $\pm$ s.e.m.

Fig. 4 | Rate of species richness change related to the rate of temperature change and precipitation change across all sampled mountains in Europe. a, Rate of species richness change (ΔSR per year $= (\text{SR}_{t2} - \text{SR}_{t1})/(t_2 - t_1)$) related to the rate of temperature change. **b**, Rate of species richness change related to the rate of precipitation change. Note that this pattern differs considerably among regions (see Extended Data Fig. 2 for more details at the regional level). Dots are semi-transparent, with darker symbols indicating overlapping points. Trend lines and R^2 values are based on univariate linear regressions and significance, indicated by stars, is based on F statistics (see Methods and Extended Data Table 3). The relationship between change in species richness and accumulated nitrogen (not shown) is not significant because nitrogen deposition varies strongly across Europe whereas the change in species richness shows the same trend across the continent. Figures and models are based on 396 observations (comparison of all 698 survey and resurveys for the 302 summits). See text and Methods for more detailed analyses with generalized mixed effects offset models, including regional differences.

of locally observed upward shifts of species ranges in mountains[12,13,20] (Fig. 2) and their recent acceleration[16,21]. Our findings thus align with those of shorter-term studies demonstrating plant community thermo-philization[15,17] and range shifts driven by warming[7].

The observed relationship between temperature change and species richness change over the past 145 years is consistent across all nine regions. Changes in precipitation and nitrogen deposition also correlate regionally with changes in species richness, but the direction and magnitude of these effects differ strongly among regions (Extended Data Fig. 2). Although precipitation change (ΔP per year) has a moderate (positive) effect on species richness trends across Europe (Extended Data Table 3, Fig. 4b), its effect is not consistent and significant across all analysed regions (Extended Data Table 4, Extended Data Fig. 2) and is minor compared to the effect of temperature change (ΔT per year; Extended Data Tables 4, 5). Changes in grazing and tourism could also affect changes in plant species richness on summits[21]. Local studies have suggested that grazing[22] and frequent disturbance by tourists[15] may suppress the elevational advance of alpine plants in response to warming in mountains. Although quantification of these relationships is challenging, locally declining levels of domestic livestock have often coincided with recovery of wild ungulate populations. Hiking tourism has increased on some summits, but intensities of human impact vary strongly. Land-use changes may thus explain parts of the local variation in species richness trends, but they vary greatly within and among regions. Without a consistent impact on species re-distribution, it is unlikely that changes in grazing and tourism can account for the consistent, continent-wide increase in plant species richness evident in our data.

Some previous observations have suggested that upslope species migration in mountains occurs almost in synchrony with climate warming[17], whereas findings from other studies indicate that long lags in dispersal, establishment and extinction can be expected for many alpine plant species[23,24]. We systematically tested for time-lags (up to 10 years) in increases in species richness following changes in climate, but found that the inclusion of time-lags did not significantly improve the explanatory power of our models (Extended Data Table 6). This finding suggests that increases in species richness on European summits are a direct and immediate response to climate warming (Fig. 2) and, thus, can be expected to accelerate further as climate warming continues
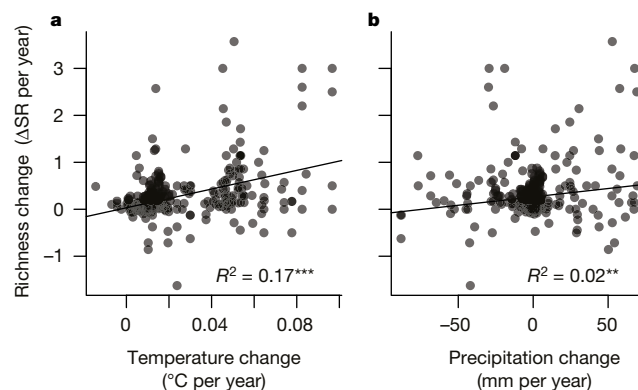
to accelerate[1]. However, because we focus on the average trend and do not account for non-colonizing lower-altitude species, we cannot exclude the possibility that only a fraction of species responded quickly to climate change, thus creating the observed relationship, while an unknown number of species lags behind the change in climate. Our observations may, therefore, underestimate the expected long-term species turnover on summits.

The accelerated increase in species richness on mountain summits is likely to result from an upward shift in the upper range limits of an increasing number of species. Trait analyses show that new colonizers exhibit growth strategies characteristic of species from lower elevations, such as larger size ($P < 0.001$), higher specific leaf area ($P < 0.001$) and a general association with warmer temperatures ($P < 0.001$; Extended Data Table 7) compared to established species. Ultimately, the lower range limits of species will also shift upwards, but these limits are often determined and changed by biotic interactions and are, therefore, only indirectly related to temperature[25]. As more species become established at high-elevation sites, local extinctions will be likely to result from competitive replacement of slow-growing, stress-tolerant alpine species by more vigorous generalists that benefit from warming, rather than by direct adverse effects of warming on the summit species[26]. However, competitive replacement of resident species requires that colonizers build up sufficiently large populations. Local extinctions should hence follow colonization with a time-lag. Consequently, accelerating plant species richness increases are expected to be a transient phenomenon that hides the accumulation of a so-called extinction debt[23,27]. The relaxation time until this debt is paid off is likely to be characterized by continuous shifts in abundance ratios, which may serve as sensitive early warning signals of upcoming extinctions[15]. The length of this relaxation time will probably depend on factors such as the longevity of high-elevation species, plant clonal abilities and the local microhabitat diversity, supporting the persistence of cold-climate microrefugia for high-alpine species[28,29]. Although these processes, along with species' intrinsic ability to tolerate changing climates, may buffer local extinctions, a rapid loss of alpine-nival species may occur

under accelerated climate warming. Additionally, if major changes and extinctions in alpine systems are not gradual, but are instead initiated by threshold-like dynamics (for example, shrub and tree encroachment), critical tipping points may be approached with increasing speed under accelerated climate warming.

Our results underline the link between accelerating climate warming and species richness change in mountains. We thus provide a particularly compelling example of the human-driven impact on terrestrial biota that is highly consistent with the recently reported Great Acceleration in Earth system trends in the Anthropocene and strikingly synchronous with the recent accelerating trends observed in many socio-economic indicators[6]. The observed acceleration of biodiversity change in mountain ecosystems highlights the rapid and widespread consequences of human activities on the biosphere, with important consequences for ecosystem functioning, human wellbeing, and the dynamics of climate change[30].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0005-6.

1. Smith, S. J., Edmonds, J., Hartin, C. A., Mundra, A. & Calvin, K. Near-term acceleration in the rate of temperature change. *Nat. Clim. Chang* **5**, 333–336 (2015).
2. Comiso, J. C., Parkinson, C. L., Gersten, R. & Stock, L. Accelerated decline in the Arctic sea ice cover. *Geophys. Res. Lett* **35**, L01703 (2008).
3. Kintisch, E. Sea ice retreat said to accelerate Greenland melting. *Science* **352**, 1377 (2016).
4. Hughes, T. P. et al. Global warming and recurrent mass bleaching of corals. *Nature* **543**, 373–377 (2017).
5. Hollesen, J., Matthiesen, H., Møller, A. B. & Elberling, B. Permafrost thawing in organic Arctic soils accelerated by ground heat production. *Nat. Clim. Chang* **5**, 574–578 (2015).
6. Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: The Great Acceleration. *Anthropocene Rev* **2**, 81–98 (2015).
7. Alstad, A. O. et al. The pace of plant community change is accelerating in remnant prairies. *Sci. Adv* **2**, e1500975 (2016).
8. Chen, I. C., Hill, J. K., Ohlemüller, R., Roy, D. B. & Thomas, C. D. Rapid range shifts of species associated with high levels of climate warming. *Science* **333**, 1024–1026 (2011).
9. Wiens, J. J. Climate-related local extinctions are already widespread among plant and animal species. *PLoS Biol* **14**, e2001104 (2016).
10. Gobiet, A. et al. 21st century climate change in the European Alps—a review. *Sci. Total Environ* **493**, 1138–1151 (2014).
11. Mountain Research Initiative EDW Working Group Elevation-dependent warming in mountain regions of the world. *Nat. Clim. Chang.* **5**, 424–430 (2015).
12. Lenoir, J., Gégout, J.-C., Marquet, P. A., de Ruffray, P. & Brisse, H. A significant upward shift in plant species optimum elevation during the 20th century. *Science* **320**, 1768–1771 (2008).
13. Pauli, H. et al. Recent plant diversity changes on Europe's mountain summits. *Science* **336**, 353–355 (2012).
14. Grytnes, J.-A. et al. Identifying driving factors behind observed species range shifts on European mountains. *Glob. Ecol. Biogeogr* **23**, 876–884 (2014).
15. Gottfried, M. et al. Continent-wide response of mountain vegetation to climate change. *Nat. Clim. Chang* **2**, 111–115 (2012).
16. Wipf, S., Stöckli, V., Herz, K. & Rixen, C. The oldest monitoring site of the Alps revisited: Accelerated increase in plant species richness on Piz Linard summit since 1835. *Plant Ecol. Divers* **6**, 447–455 (2013).
17. Bertrand, R. et al. Changes in plant community composition lag behind climate warming in lowland forests. *Nature* **479**, 517–520 (2011).
18. Stöckli, V., Wipf, S., Nilsson, C. & Rixen, C. Using historical plant surveys to track biodiversity on mountain summits. *Plant Ecol. Divers* **4**, 415–425 (2012).
19. Verheyen, K. et al. Combining biodiversity resurveys across regions to advance global change research. *Bioscience* **67**, 73–83 (2017).
20. Odland, A., Høitomt, T. & Olsen, S. L. Increasing vascular plant richness on 13 high mountain summits in Southern Norway since the early 1970s. *Arct. Antarct. Alp. Res* **42**, 458–470 (2010).
21. Walther, G.-R., Beißner, S. & Burga, C. A. Trends in the upward shift of alpine plants. *J. Veg. Sci* **16**, 541–548 (2005).
22. Speed, J. D. M., Austrheim, G., Hester, A. J. & Mysterud, A. Elevational advance of alpine plant communities is buffered by herbivory. *J. Veg. Sci* **23**, 617–625 (2012).
23. Dullinger, S. et al. Extinction debt of high-mountain plants under twenty-first-century climate change. *Nat. Clim. Chang* **2**, 619–622 (2012).
24. Hülber, K. et al. Uncertainty in predicting range dynamics of endemic alpine plants under climate warming. *Glob. Change Biol* **22**, 2608–2619 (2016).
25. Vetaas, O. R. Realized and potential climate niches: a comparison of four *Rhododendron* tree species. *J. Biogeogr* **29**, 545–554 (2002).
26. Alexander, J. M., Diez, J. M. & Levine, J. M. Novel competitors shape species' responses to climate change. *Nature* **525**, 515–518 (2015).
27. Cotto, O. et al. A dynamic eco-evolutionary model predicts slow response of alpine plants to climate warming. *Nat. Commun* **8**, 15399 (2017).
28. Kulonen, A., Imboden, R. A., Rixen, C., Maier, S. B. & Wipf, S. Enough space in a warmer world? Microhabitat diversity and small-scale distribution of alpine plants on mountain summits. *Divers. Distrib* **24**, 252–261 (2018).
29. Winkler, M. et al. The rich sides of mountain summits — a pan-European view on aspect preferences of alpine plants. *J. Biogeogr* **43**, 2261–2273 (2016).
30. Pecl, G. T. et al. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* **355**, eaai9214 (2017).
31. Zemanek, A. Bogumił Pawłowski (1898–1971) — życie i dzieło. *Fragm. Florist. Geobot. Polon* **19**, 205–244 (2012).
32. Ellenberg, H. J. Braun-Blanquet 3.8.1884–22.9.1980 R. Tüxen 21.5.1899–16.5.1980–Jahre Pflanzensoziologie. *Ber. Deutsch. Bot. Ges* **95**, 387–391 (1982).
33. Unknown. Tschechoslowakei 1928. Mohelno. Prof. Rübel im Jihlavka-Tale. ETH-Bibliothek Zürich, Bildarchiv (1928).

Author contributions S.W., M.J.S., J.-A.G., G.J., A.K., C.R., A.D.B., S.H., J.K., K.K., I.K., J.L., S.N., H.P., P.V. and M.W. elaborated the concept; A.K., S.W., M.B.-D., H.P., C.R., P.V. and M.W. organized and harmonized data; M.J.S. implemented the data analyses with support from other authors, particularly S.W., J.-A.G., J.L., J.-C.S., S.D. and D.G.; E.B., S.B., F.T.B., P.C., M.A.D., A.D., B.E., V.A.F., O.F.-A., K.F.F., D.G.-G., E.T.G., J.-A.G., S.V.H., H.H., M.J.H., B.J., F.J., R.K., K.K., J.K., A.L., M.M., U.M.d.C., A.O., S.L.O., S.P., H.P., M.P., V.P., B.S., K.S., V.S., C.R., G.T., J.-P.T., P.V., S.J.W., S.W and N.E.Z. contributed data. M.J.S. led the manuscript writing, with contributions from all authors.

METHODS

Vegetation resurveys on European mountain summits. Precise relocation of vegetation records is possible on mountain summits. European botanists, fascinated by the limits of plant life, noted this potential more than a century ago (Fig. 1)[18]: "On the basis of a comprehensive description of locations, it will not be difficult to verify my species lists, and an increase or decrease of species richness in the future will be possible to detect with high certainty," (Josias Braun-Blanquet in 1913, translated from[34], p. 329). This foresight and the data these botanists gathered on mountain summits give us the opportunity to study the effect of accelerated warming on plant species richness. Thus, summits are optimal for resurveys of species occurrences and for detecting changes in plant species richness over time, even when the first surveys were carried out before the GPS era. In this study, 302 summits with historical vegetation records were resurveyed between one and six times, resulting in a total of $n = 698$ surveys. All vegetation surveys were conducted in summer. For each survey, all plant species occurring on the summit (generally delineated by the uppermost 10 m of elevation)[35] were noted. Vegetation surveys were compared for each summit. Species names were standardized to the nomenclature of Flora Europaea (or local flora for species absent in the Flora; see Supplementary Information).

Environmental data. For each summit, mean monthly temperature and precipitation were calculated following the established change factor methodology[36], which combines statistical downscaling with temporal trend analyses. First, temporal data available from CRU TS 3.23 (0.5° resolution, 1901–2015)[37] and the European Gridded Monthly Temperature (0.5° resolution, 1765–2000)[38] were statistically related to the higher spatial resolution of WorldClim monthly mean climatic grids (30 arcsec resolution) for the overlapping period of 1950 to 2000 using the change factor method[36]. We assumed that anomalies (compared with mean value over the period 1950–2000 of the coarse-grained climatic conditions minus the climatic conditions within each smaller pixel of WorldClim) computed for the overlapping period (1950–2000) remained the same before 1950 and after 2000. Second, elevational differences between summits and the mean elevation of the corresponding WorldClim digital elevation model were included as an additional correction term ($-0.006\,°C \times \Delta$elevation (m)) for mean temperature data. By combining the two corrections, temporal trends available from the 0.5° degree resolution temporal data were corrected for differences originating from scale and climate model, and the precise elevation of the summit (temperature only). While we consider the resulting temporal trends for the temperature data to be reliable owing to the generally higher spatial and temporal autocorrelation and a higher correlation with elevation, the precipitation data do not show a systematic change with elevation and are less predictable over small spatial distances[39] and, therefore, need to be interpreted more cautiously. Environmental variables were included in the models after calculating temporal changes (see 'Importance of environmental drivers'). Consequently, environmental variables are unbiased by weaknesses in the spatial interpolations. For temperature and precipitation, time series from CRU TS 3.23 (1901–2015) and the European Gridded Monthly Temperature (1765–2000) were combined to match the study period (1880–2016) by taking the mean per grid cell for the overlapping years (Spearman $r = 0.97$ for the overlapping period 1901–2000). As neither of the two data sources extends to 2016, climate values for 2015 were taken again for 2016 for the 19 affected summits. Furthermore, historical nitrogen deposition data (NH_x and NO_x modelled from 1850 to 2010) were extracted from the European Fluxes Database (http://www.europe-fluxdata.eu/) and extrapolated for the missing five years (2011–2016). The data originate from the global chemistry transport model version 5 (TM5, annual data with a 0.25° latitude/longitude resolution)[40]. Data handling and all subsequent analyses were conducted in R version 3.3.1[41].

The velocity of species richness changes. Species richness (SR) on mountain summits was analysed for its change with time (t, year of record) across all summits by implementing a generalized linear mixed effects model (GLMM) with a Poisson family error distribution (SR $\sim t$) and a random effect (intercept) of mountain to account for repeated samples (GLMM 1 in Extended Data Table 1; all mixed effects models were built with R package *lme4* version 1.1-12)[42]. Further, we ran the models including random effects (intercept) of region (mountain nested in region; GLMM 2 in Extended Data Table 1) and observation ID (to account for overdispersion[43]; GLMM 3 in Extended Data Table 1). All models provided qualitatively equivalent results (Extended Data Table 1). We repeated all GLMMs allowing a breakpoint (bp) in the relationship between species richness and time by fitting independent slope coefficients for the time period before and after the breakpoint (SR $\sim$ ifelse($t < $ bp, bp $- t$, 0) $+$ ifelse($t < $ bp, 0, $t - $ bp) $+$ random structure). The breakpoint was fitted independently by minimizing the model deviance (Extended Data Table 1).

Acceleration of species richness changes. The potential acceleration in the average velocity of species richness changes on mountain summits between 1871 and 2016 was tested by means of a linear mixed effects model (LMM) with a Gaussian family error distribution (ΔSR/$\Delta t \sim t_{MP}$). With the model, we analysed the rate of change

in species richness over time (midpoint year between two surveys $t_{MP} = (t_1 + t_2)/2$). The dependent variable ΔSR/Δt was calculated from the difference in species richness and the difference between years of observation of two consecutive surveys on the same summit ($(SR_{t2} - SR_{t1})/(t_2 - t_1)$). A random effect (intercept) of mountain was included to account for repeated samples. We also ran the model including a random effect (intercept) of mountain nested within region but found qualitatively similar results (Extended Data Table 2). Mathematically, ΔSR/Δt is independent of richness on the summits as well as of time elapsed between sequential visits on the summit. However, more species-rich summits seemed to be associated with higher rates of change, as indicated by a significant positive effect if the species richness of the first survey was included as an explanatory variable in the fixed component of the LMM (Extended Data Table 2). We also tested whether there was an effect of the number of years between two consecutive surveys on ΔSR/Δt, as a longer period between surveys might mask short-term fluctuations, but this effect was not significant (Extended Data Table 2).

A linear increase in the rate of change with time (ΔSR/$\Delta t \sim t_{MP}$) corresponds to an accelerated richness increase. As Figs. 2 and 3 indicate a nonlinearity in the relationship, we also ran all models allowing a breakpoint in the relationship between the rate of change and the time between surveys (Extended Data Table 2). It is likely that the real breakpoint (compared with the onset) of the acceleration trend in the increase in plant species richness happened slightly later than the breakpoint suggested by this particular analysis. Indeed, the estimated breakpoint approximates the timing of change as the year between two sequential surveys and thus mechanistically moves every change temporally towards the median of the time series.

In the raw data, the average rate of species richness increase per summit was found to be much higher in the past decade (2007–2016; $+2.9$ species) compared to fifty years earlier (1957–1966, $+1.1$ species). When the slopes are averaged across all summits with an observation before and after a given year, inversely weighted by the number of years between observations (to account for temporal resolution, as a longer period between surveys might mask short-term fluctuations), the differences become even more apparent ($+5.4$ species in the past decade as opposed to $+1.1$ species per decade fifty years earlier).

We analysed changes in absolute species numbers, as relative changes are sensitive to the richness values to which they are normalized. Still, repeating the linear mixed effects model with the changes in relative species richness (calculated by taking the difference between survey and resurvey normalized by resurvey richness and years between observations) revealed equivalent results and the same conclusions as using changes in absolute species numbers over time.

Visualization of temporal changes in richness. The average richness change per year (ΔSR/$\Delta t = (SR_{t2} - SR_{t1})/(t_2 - t_1)$) across all summits was calculated (Extended Data Fig. 1a). Figure 3 displays how the average in ΔSR/Δt across all summits changed over time. As values for ΔSR/Δt originating from summits with a higher temporal sampling density better represent the instant rate of change for that specific year (t), we inversely weighted the calculated values for ΔSR/Δt by the difference in years between observations ($t_2 - t_1$) to account for temporal resolution.

The changes in species richness per year (ΔSR/Δt) accumulate over time and result in an absolute change in species richness (Extended Data Fig. 1b). These absolute changes in species richness are visualized for each region in Fig. 2 (black line). In order also to visualize variability within regions, confidence intervals were calculated on the basis of the standard deviation of richness change among summits in a region (Extended Data Fig. 1c, d).

Importance of environmental drivers. The average velocity of species richness changes (ΔSR/Δt) was related to the change in mean annual temperature ($\Delta T/\Delta t$; T is temperature) and precipitation ($\Delta P/\Delta t$; P is precipitation) for the same period (see below for further details), as well as to the accumulated nitrogen deposition (N_{accum}, details explained below) across all summits, by implementing LMMs with a Gaussian family error distribution that included each of the three potential explanatory variables (different rows in Extended Data Table 3, model formula can be seen in table caption). Variable performance was compared using the corrected version (for small sample size) of the Akaike Information Criteria (AICc)[44]. All LMMs consistently detected a clear positive relationship between species richness changes and temperature changes, while a slightly weaker positive relationship with precipitation changes was detected. In particular, the relationship with temperature change is surprisingly strong considering that climate models are built on long-term air temperature measurements at two metres above ground in climate stations that are mainly located in valleys and can only approximate changes in growth conditions for summits species. No relationship with the accumulated nitrogen deposition was detected across Europe (Extended Data Table 3).

The explanatory variables $\Delta T/\Delta t$ and $\Delta P/\Delta t$ were calculated as the mean change per year (for example, $\Delta T/\Delta t = (T_{t2} - T_{t1})/(t_2 - t_1)$). Climate variables such as temperature and precipitation are usually integrated over longer time periods to level out short-term fluctuations. As we were interested in the effect of such shorter-term fluctuations, we systematically tested which periods would

provide the best fit within our LMM framework (1–30 years). Besides mean annual temperature and precipitation, we further tested alternative measurements of the climate variables. If species' ranges were limited primarily by growing season temperatures, we would expect spring and summer warming to best explain temporal changes in species richness. Alternatively, if many alpine species were limited not by growing season temperature, but rather by climatic extremes, winter temperatures or precipitation might be more important in determining which species can survive in a given location. We therefore systematically pre-analysed temperature and precipitation variables by testing for the effect of winter precipitation (December–February) and of snow accumulation (precipitation in months with a mean temperature below freezing).

Further, nitrogen from deposition may accumulate in the soil, particularly in high elevation systems with limited resource cycling[45,46]. In our data, nitrogen deposition has declined sharply in recent decades[39], although its accumulated effect may still influence community dynamics[47]. We thus calculated accumulated deposition of both NH_4 and NO_3 since 1850 for each vegetation survey.

The systematic test of different variables and time periods (Extended Data Table 5) identified annual summer temperature (15-year mean), annual precipitation (1-year mean) and NO_3 (referred to as N_{accum}) as the most suitable predictors, and these variables were then used in all subsequent analyses. As this type of variable selection biases analyses towards significant relationships, all analyses were repeated with mean annual values (10-year mean), resulting in qualitatively similar results. Model residuals were visually checked for temporal autocorrelation, and there was no sign of a temporal trend in the residuals.

Time-lags in richness change. Biotic responses may show a delayed response to climate change[17,24], as species may need considerable time to spread and establish (compare migration and establishment lags). Therefore, observed species richness on a mountain summit at a given point in time may reflect climatic conditions from several years earlier. A systematic time-lag was therefore implemented between our species observations and the climate period used to relate the average velocity of species richness changes to changes in climatic conditions, and an increase in explanatory power by including a time-lag (5 or 10 years) was tested (Extended Data Table 6). Final results are presented without time-lags because including them did not increase the power in our analyses to explain the average velocity of species richness changes.

An alternative approach to analysing the average velocity of species richness changes ($\Delta SR/\Delta t$) with rates of change in environmental predictors ($\Delta T/\Delta t$, $\Delta P/\Delta t$; see Extended Data Table 3) is to directly relate species richness changes (ΔSR) to changes in environmental variables over the same period (ΔT, ΔP). This approach is more intuitive (and closer to the data) but ignores differences in time between sampling events. Analyses using this approach yielded results qualitatively similar to the results of the main analysis (Extended Data Table 3), with the exception that the effect of precipitation changes was not significant (Extended Data Table 4).

Trait-based analyses. *Differing trait signal in colonizing species.* Changes in plant life strategies and dispersal constraints would be represented by a systematic difference in indicative traits. We thus compared specific leaf area (SLA)[48], plant height[48] and seed mass[48] among colonizing species and species in the resident community, using a LMM framework with 'resurvey' as a random effect. To test for the colonization and establishment, within the recipient community, of warmth-tolerating species from lower elevations, we used Landolt species indicator values for temperature[49]. For 364 resurveys (12,738 observations for 873 species), direct comparisons of plant trait values of newly established colonizers (that is, additional species recorded in a resurvey) with those of species that had been present in the previous survey (recipient community) indicate significantly increased SLA ($P < 0.001$) and plant height ($P < 0.001$) of successful colonizers, but no significant difference in seed mass ($P = 0.85$). Colonizers were also more adapted to warmer climates (showing higher Landolt temperature values) than species of the resident community ($P < 0.001$; Extended Data Table 7).

Data reliability. *Sampling intensity.* Our analysis of the rate of change is relatively robust with respect to different sampling periods. The increasing sampling frequency over time (Fig. 1) helped to reliably quantify the rates of change in later time periods and thus to support our conclusion of an acceleration in richness change. Consistent continent-wide and short-term fluctuations in species richness that might have occurred in the early 20th century would be likely to go undetected owing to the low data availability in the early 20th century of our time series data, but long-term trends would be clearly visible. There is, however, no evidence that the unbalanced sampling effort over time and different sampling intervals hide unobserved fluctuations in early periods. In line with this, the summits for which we have a large number of repeated surveys show small short-term fluctuations but confirm the detected steady increase of richness over time and an acceleration in recent years[16].

Observer errors. Previous studies explicitly addressing observer errors in summit resurveys have demonstrated reliable quantification of vegetation change over long time periods[35]. Many of the early records were collected by expert botanists with a scientific interest in long-term changes and the explicit aim of enabling accurate later resurveys. To further reduce potential sampling and observer errors, recent resurveys were conducted without knowledge of the past species lists because surveyors who know the historical species composition have a higher chance of finding certain species again. Approximately 15% of all summits of this dataset have species records collected in the 1980s and 1990s (they were even carried out by the same people in some cases). Even if these early re-surveyors also considered the above methodological issues, we cannot rule out that the observer effort of the early re-surveyors was greater than that of the historical surveyors. However, our carefully implemented re-survey methodology made sure that our recent observer effort did not exceed that of the early re-surveyors during the 1980s and 1990s. Given this, the clear signal that most of the increase in species richness occurred after the 1980s and 1990s is a strong indication that a possible increase in observer effort, if present, is responsible for only a limited amount of the increase in species richness. We are, thus, confident that observer errors did not systematically influence our analyses.

Summit area. Summit area may affect the observed changes in species richness, probably through its effect on species richness (compare with species–area relationship). We cover this potential effect of area on the change in species richness by including absolute species richness as a co-variate in our analyses. A potential direct effect of area could be tested only for the summits within Switzerland, as data of sufficient spatial resolution to calculate the surface of the uppermost 10 m of a summit was available to us only from this country (swissALTI3D model, a digital elevation model with 2-m resolution). The summit area for Swiss summits varied by as much as 40 times (392–16,720 m^2). Surprisingly, regression analyses indicated that there was no significant effect of summit area on the historical or recent species number, or on the change in species number (area was log-transformed to reach normal distribution). Further evidence of a limited effect of summit area is indicated by the fact that in recent resurveys the species numbers of historic surveys were reached within the uppermost 4–5 m of each summit, which on a summit with conical shape corresponds to a much smaller area than the originally sampled uppermost 10 m (Extended Data Fig. 3). We conclude that, on mountain summits, factors independent of area, for example, environmental conditions and micro-topographic variability[28], seem much more important for species richness or changes thereof than area per se.

Data and code availability. Data and R code are available from the corresponding authors.

34. Braun, J. Die Vegetationsverhältnisse der Schneestufe in den Rätisch-Lepontischen Alpen. Ein Bild des Pflanzenlebens an seinen äußersten Grenzen. *Neue Denkschr. Schweiz. Naturf. Ges* **48**, 1–347 (1913).
35. Burg, S., Rixen, C., Stöckli, V. & Wipf, S. Observation bias and its causes in botanical surveys on high-alpine summits. *J. Veg. Sci* **26**, 191–200 (2015).
36. Anandhi, A. et al. Examination of change factor methodologies or climate impact assessment. *Wat. Resour. Res* **47**, W03501 (2011).
37. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol* **34**, 623–642 (2014).
38. Casty, C., Raible, C. C., Stocker, T. F., Wanner, H. & Luterbacher, J. *European Gridded Monthly Temperature, Precipitation and 500hPa Reconstructions*; IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2008-023 (NOAA/NCDC Paleoclimatology Program, Boulder, 2008).
39. Daly, C., Neilson, R. P. & Phillips, D. L. A statistical topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**, 140–158 (1994).
40. Huijnen, V. et al. The global chemistry transport model TM5: description and evaluation of the tropospheric chemistry version 3.0. *Geosci. Model Dev* **3**, 445–473 (2010).
41. R Core Team R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, 2016).
42. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw* **67**, 1–48 (2015).
43. Harrison, X. A. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* **2**, e616 (2014).
44. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* 2nd edn (Springer, New York, 2002).
45. Burns, D. A. The effects of atmospheric nitrogen deposition in the Rocky Mountains of Colorado and southern Wyoming, USA—a critical review. *Environ. Pollut.* **127**, 257–269 (2004).
46. Körner, C. Mountain ecosystems in a changing environment. *Ecomont* **6**, 71–77 (2014).
47. Bobbink, R. et al. Global assessment of nitrogen deposition effects on terrestrial plant diversity: a synthesis. *Ecol. Appl.* **20**, 30–59 (2010).
48. Kleyer, M. et al. The LEDA Traitbase: a database of life-history traits of Northwest European flora. *J. Ecol* **96**, 1266–1274 (2008).
49. Landolt, E. et al. *Flora Indicativa* (Haupt, Bern, 2010).
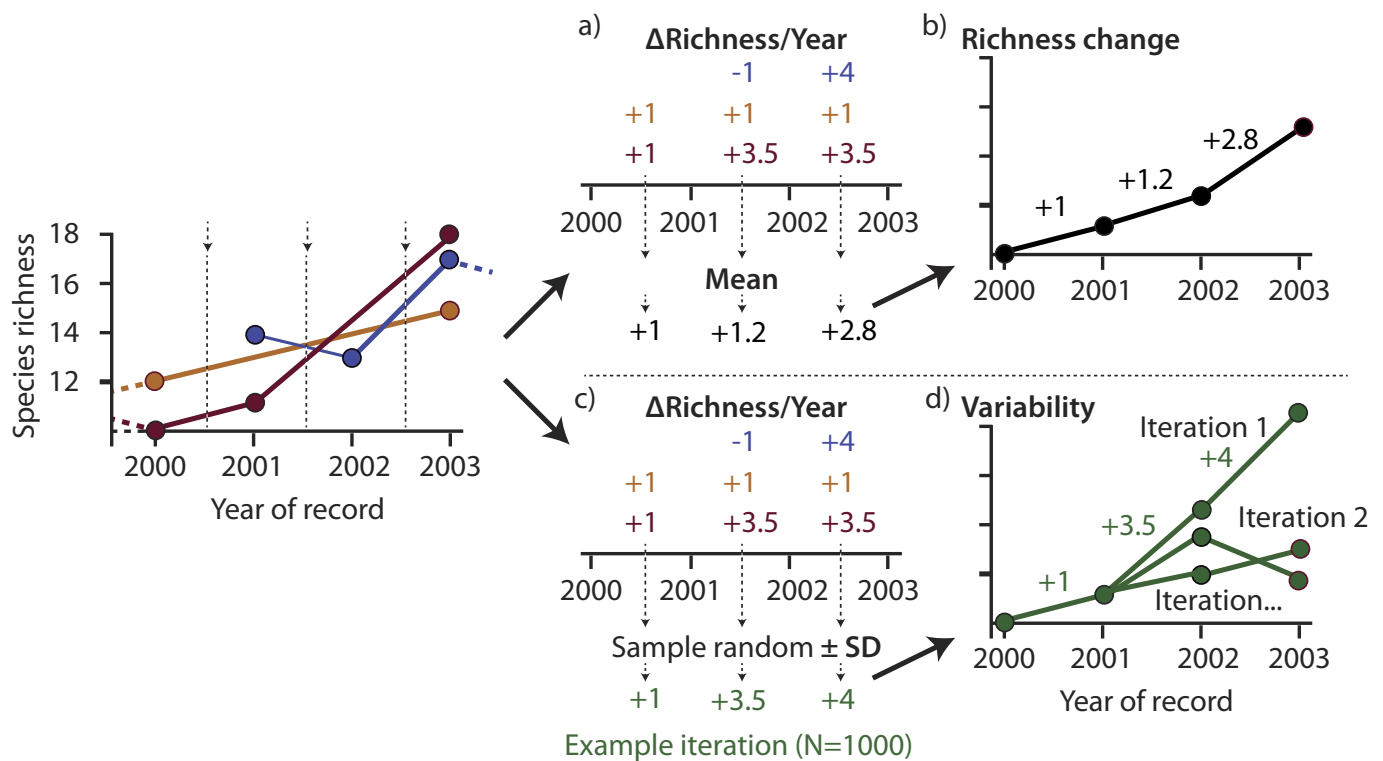
Extended Data Fig. 1 | Visualizing richness change. This conceptual figure shows the approach implemented in the main text to visualize richness change over time based on the raw data (Figs. 2, 3). **a**, The mean richness change per year ($\Delta SR/\Delta t = (SR_{t2} - SR_{t1})/(t_2 - t_1)$) across all summits was calculated (Fig. 3). **b**, The mean richness change per year accumulates with time to yield absolute changes in species richness (black line in Fig. 2). **c, d**, Variability in the absolute change in species richness was visualized by randomly sampling ΔSR from all mountains available each year, but adding the s.d. within a region and year. The displayed range in Fig. 2 illustrates the 5th and 95th percentiles of the resulting richness change values from 1,000 runs (orange shading in Fig. 2). This approach reveals changes in variability among mountains over time while also showing overall variability for time steps where only a few summits were sampled (particularly in early time periods).
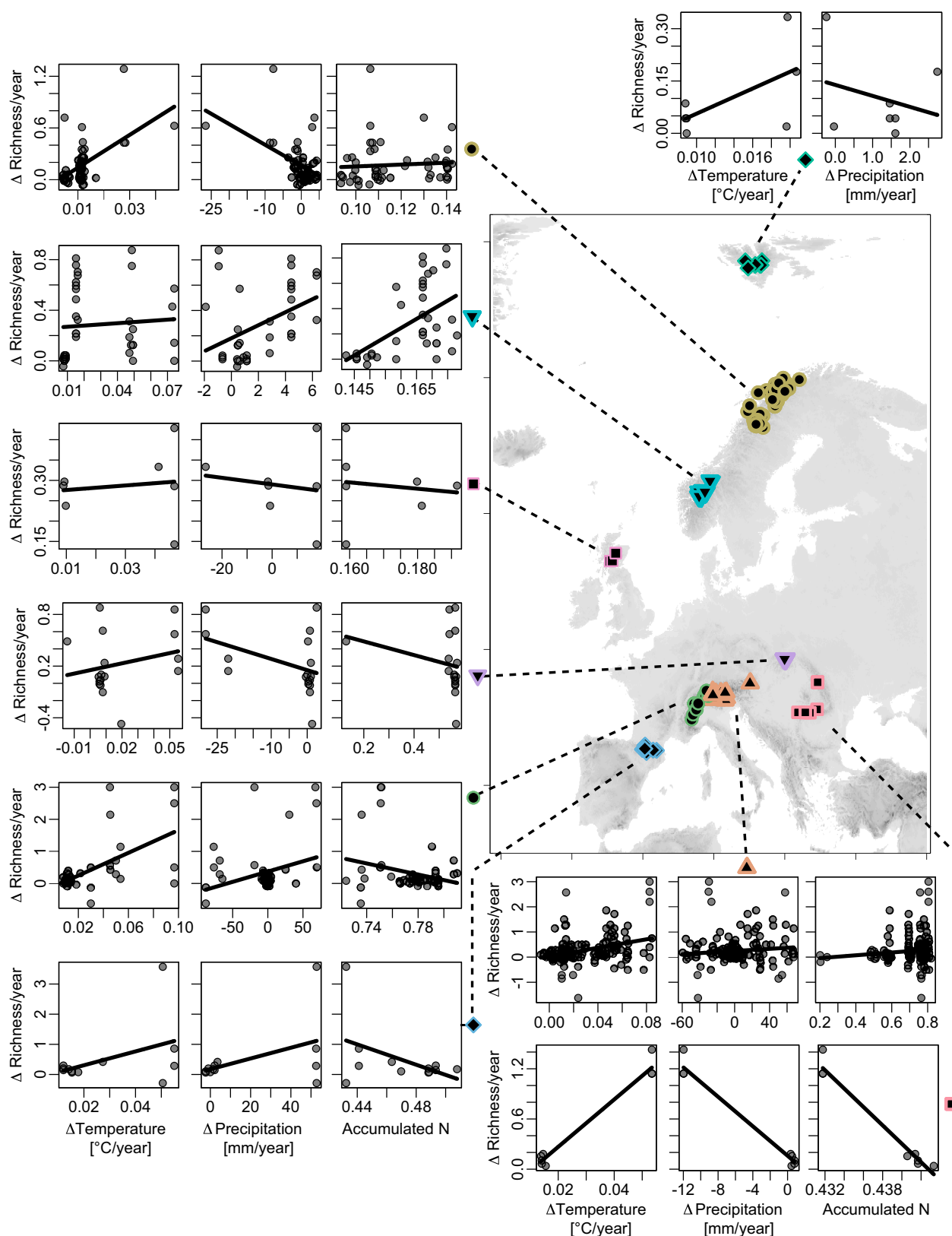
Extended Data Fig. 2 | Relationship between rates of changes in species richness across Europe and rates of increase in temperature (left column), rates of change in precipitation (middle column) and accumulated nitrogen deposition (right column). Trend lines are interpolated from a simple linear model and are in many cases not significant. Species richness was quantified as the difference between vegetation surveys from the same summit at different times (Extended Data Fig. 1). No nitrogen data were available for Svalbard. The number of observations (comparison of survey and resurveys) are: Svalbard, 7; Northern Scandes, 54; Southern Scandes, 27; Scotland, 7; NW Carpathians, 16; Eastern Alps, 122; Western Alps, 48; SE Carpathians, 9; Pyrenees, 12 (see Fig. 1 for more details).
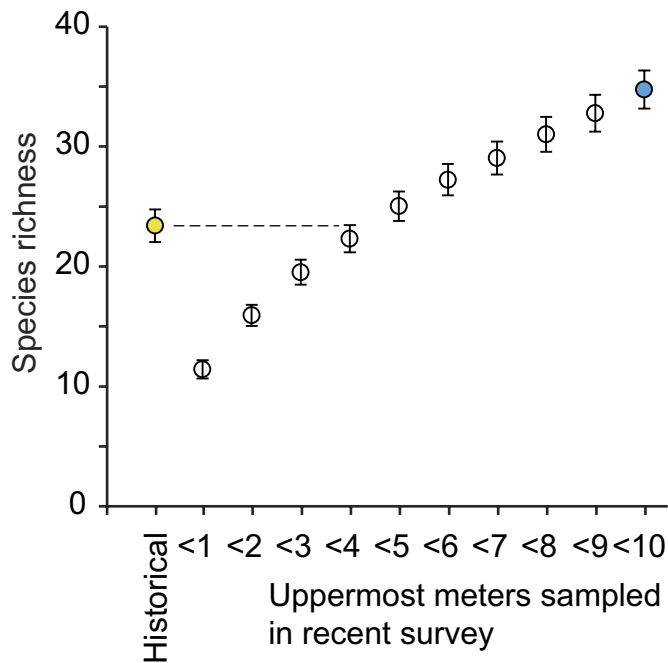
Extended Data Fig. 3 | Historical and recent species richness versus sampling area. Historical species richness was exceeded within a small sampling area during recent resurveys. Species richness of the historical survey (yellow) contrasted with a species richness accumulation curve of the recent surveys on summits where the highest occurrence of each recent species was estimated to the nearest 1-m elevation. The number of species found historically within the uppermost 10 m of a summit was exceeded within the uppermost 5 m in the most recent resurveys. This analysis includes all 157 European summits for which such data are available, regardless of whether the historical species number was reached in recent times. The blue circle visualizes average species richness of the recent surveys within the uppermost 10 m.

Extended Data Table 1 | Increase in species richness with time

	Fixed effect (coefficients ±std. error)			**Random effects** (std. deviations)			
Model	Intercept	Year of record		Mountain	Region:Mount.	ID	AICc
GLMM 1	-5.84 ± 0.35***	0.004 ±0.0002***		0.97	-	-	5758
GLMM 2	-5.84 ± 0.35***	0.004 ±0.0002***		0.88	0.41	-	5760
GLMM 3	-7.31 ± 0.57***	0.005 ±0.0003***		0.75	0.60	0.22	5585
GLM	-7.60 ± 0.33***	0.006 ±0.0002***		-	-	-	18256
Model	Intercept	Time < BP	Time > BP	Mountain	Region:Mount.	ID	AICc
GLMBM	2.73 ±0.07***	0.001 ±0.001	0.013 ±0.001***	0.96	-	-	5684
GLMBM	2.73 ±0.07***	0.001 ±0.001	0.013 ±0.001***	0.87	0.41	-	5686
GLMBM	2.64 ±0.07***	0.001 ±0.003	0.006 ±0.0004***	0.83	0.49	0.22	5583

Generalized linear mixed effects models (Poisson family error distribution) show an increase in species richness with time (richness ~ year of record). Different random effect structures were applied. The lower panel includes a breakpoint in the relationship between rate of richness change and time. The breakpoint was fitted independently by minimizing model deviance and was estimated around the year 1970. All models are based on 698 observations. Significant effects are indicated by asterisks (***$P < 0.001$). GLMM, generalized linear mixed effects model; GLM, generalized linear model; GLMBM, generalized linear mixed effects breakpoint model; ID, observation ID.

Extended Data Table 2 | Acceleration of the increase in species richness over time

Fixed effect (coefficients ±std. error)		Richness	Period	**Random effect** (std. deviations)		
Intercept	*Time*	*Richness*	*Period*	*Mountain*	*Region: Mount.*	*AICc*
-15.5±2.06***	0.008±0.001***	-	-	5.8×10⁻⁸	-	570.1
-15.5±2.06***	0.008±0.001***		-	0.0	0.0	572.1
-13.4±2.05***	0.007±0.001***	0.004±0.001***	-	0.0	-	561.7
-11.7±4.76* (p=0.014)	0.006±0.002* (p=0.012)	0.004±0.001***	n.s.	0.0	-	575.1
-13.4±2.05***	0.007±0.001***	0.004±0.001***	-	-	-	529.9

Intercept	*Time < BP*	*Time > BP*	*Richness*	*Period*	*Mountain*	*Region: Mount.*	*AICc*
0.07±0.05	0.002±0.003	0.013±0.002***	-	-	0.0	-	571.0
0.07±0.05	0.002±0.003	0.013±0.002***			0.0	0.0	573.1
0.02±0.05	0.0001±0.003	0.011±0.002***	0.004±0.001***	-	0.0	-	567.8
-0.09±0.14	0.0004±0.004	0.012±0.004***	0.004±0.001***	n.s.	0.0	-	580.7
0.02±0.05	0.0001±0.003	0.011±0.002***	0.004±0.001***	-	-	-	527.0

Linear mixed effects models (Gaussian family error distribution) showed an acceleration of the increase in species richness over time ($\Delta SR/\Delta t \sim t$). Different random effect structures were implemented. The species richness from the summit's first survey and the number of years between two consecutive observations (period) were included as additional explanatory variables. The lower panel further includes a breakpoint in the relationship between rate of richness change and time. The breakpoint was fitted independently by minimizing model deviance and was estimated for the year 1971. All models were based on 396 observations (comparison of survey and resurveys). Significant effects are indicated by asterisks (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; P values > 0.001 are additionally reported in brackets). Note that models without random structure performed best.

Extended Data Table 3 | Explanatory variables for velocity in species richness changes

Intercept	*ΔT/Δt*	*ΔP/Δt*	N_{accum}	*Richness*	*AICc*	*AICWt*
0.01 ±0.06	9.8 ±1.1***	0.005 ±0.001***	-0.16 ±0.09	0.004 ±0.001***	488.1	0.64
-0.06 ±0.04	9.5 ±1.1***	0.005 ±0.001***	-	0.004 ±0.001***	489.3	0.36
0.03 ±0.06	9.1 ±1.1***	-	-0.17 ±0.09	0.004 ±0.001***	509.2	0.00
0.14 ±0.06*	-	0.004 ±0.001***	-0.07 ±0.10	0.006 ±0.001***	556.2	0.00

Results of linear mixed effects models (Gaussian family error) showing the relationship of the average velocity in species richness changes with the change in potential explanatory variables (temperature, precipitation, nitrogen deposition). Initial species richness on the summits was added as a further independent variable and indicated that species-rich systems showed a larger net change. The implemented model formula was lmer(ΔSR/Δt ~ ΔT/Δt + ΔP/Δt + N_{accum} + richness + (1 | mountain)). Model performance was compared using AICc, which also defines the order of models, with the best one on top. In addition, significant results from tests using F statistics are indicated by asterisks (***$P < 0.001$). All values indicate model coefficients ± s.e. Rerunning the analyses after centring (subtracting the means) and scaling (dividing by s.d.) the explanatory variables indicated a larger coefficient and thus stronger effect of temperature than that of precipitation (ΔSR/Δt = 0.00 (±0.04) + 0.39 (±0.05) × ΔT/Δt*** + 0.22 (±0.04) × ΔP/Δt*** + 0.21 (±0.05) × richness***; asterisks indicate significant effects with ***$P < 0.001$). As no nitrogen data were available for the seven summits on Svalbard, the analyses presented in the table were performed on a subset of 389 temporal comparisons (comparing surveys and resurveys resulting from 684 observations). To account for spatial autocorrelation, we further repeated the full model averaging over all summits sampled over the same time period and falling in the same grid cell of the original climate data. The results of this model were qualitatively similar (ΔSR/Δt = −0.004 (±0.05) + 9.7 (±1.1) × ΔT/Δt*** + 0.005 (±0.001) × ΔP/Δt*** − 0.14 (±0.09) × N_{accum} + 0.005 (±0.001) × richness***).

Extended Data Table 4 | Explanatory variables for species richness changes

Intercept	$\Delta T/\Delta t$	$\Delta P/\Delta t$	N_{accum}	Richness	AICc	AICWt
7.7 ±1.6***	5.8 ±1.2***	-	-5.4 ±2.3* (p= 0.02)	-	2950.2	0.56
7.3 ±1.7***	6.2 ±1.3***	0.002 ±0.002	-5.1 ±2.4* (p= 0.03)	-	2951.1	0.34
4.7 ±1.2***	5.9 ±1.3***	0.003 ±0.002	-	-	2953.6	0.10
12.0 ±1.5***	-	-0.001 ±0.002	-3.8 ±2.4	-	2969.9	0.00

Linear mixed effects models (Gaussian family error distribution) showing the direct relationship between species richness changes and changes in potential explanatory variables (temperature, precipitation, nitrogen deposition). Initial species richness on the summit was not added as a further independent variable as it did not show significant effects in any of the models. The implemented model formula was lmer(ΔSR ~ $\Delta T + \Delta P + N_{accum} +$ richness $+$ (1|mountain)). Variable performance was compared using AICc, which also sets the order of models, with the best one on top. Additional significance tests using F statistics are indicated by asterisks (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; P values > 0.001 are additionally reported in brackets). All values indicate model coefficients $\pm$ s.e. Rerunning the analyses after centring (subtracting the means) and scaling (dividing by standard deviations) indicated a larger coefficient and thus stronger effect of temperature compared to that of precipitation (ΔSR ~ 0.05 ($\pm$0.06) $+$ 0.25 ($\pm$0.05) $\times \Delta T$*** $+$ 0.05 ($\pm$0.05) $\times \Delta P - 0.11$ ($\pm$0.05) $\times N_{accum}$*). The analyses were performed with the same data as specified in Extended Data Table 3.

Extended Data Table 5 | Model evaluation for different explanatory variables and time periods

Temperature

Explanatory variable	Period	AICc	ΔAICc	AICc weights
Summer temperature	15	488.1	0.0	1.0
Annual temperature	15	496.5	8.4	0.0
Spring temperature	10	507.6	19.5	0.0
Annual temperature	7	509.1	20.9	0.0
Spring temperature	7	513.2	25.1	0.0
Summer temperature	7	514.7	26.6	0.0
Annual temperature	10	516.0	27.9	0.0
Annual temperature	30	517.4	29.3	0.0
Spring temperature	15	517.6	29.4	0.0
Summer temperature	5	526.7	38.5	0.0
Annual temperature	3	526.9	38.7	0.0
Spring temperature	30	528.3	40.2	0.0
Summer temperature	1	530.6	42.5	0.0
Summer temperature	30	532.5	44.4	0.0
Annual temperature	1	534.9	46.8	0.0
Annual temperature	5	535.5	47.3	0.0
Summer temperature	10	545.6	57.5	0.0
Spring temperature	5	546.2	58.1	0.0
Summer temperature	3	547.1	58.9	0.0
Spring temperature	1	548.2	60.1	0.0
Spring temperature	3	551.4	63.3	0.0

Precipitation

Explanatory variable	Period	AICc	ΔAICc	AICc weights
Annual precipitation	1	488.1	0.0	1.0
Snow precipitation	1	495.2	7.1	0.0
Winter precipitation	15	501.3	13.2	0.0
Annual precipitation	30	502.7	14.5	0.0
Snow precipitation	3	502.9	14.8	0.0
Winter precipitation	1	504.4	16.2	0.0
Snow precipitation	30	504.7	16.5	0.0
Winter precipitation	5	505.6	17.5	0.0
Summer precipitation	30	506.0	17.9	0.0
Winter precipitation	30	507.7	19.6	0.0
Summer precipitation	5	507.7	19.6	0.0
Snow precipitation	10	508.2	20.1	0.0
Snow precipitation	15	509.2	21.1	0.0
Snow precipitation	5	509.5	21.3	0.0
Annual precipitation	5	509.7	21.6	0.0
Annual precipitation	15	509.7	21.6	0.0
Winter precipitation	3	509.8	21.6	0.0
Annual precipitation	10	510.3	22.1	0.0
Summer precipitation	15	510.4	22.2	0.0
Summer precipitation	10	510.4	22.3	0.0
Summer precipitation	3	510.6	22.5	0.0
Summer precipitation	7	510.9	22.8	0.0
Winter precipitation	10	511.0	22.9	0.0
Annual precipitation	3	511.2	23.1	0.0
Annual precipitation	7	511.2	23.1	0.0
Snow precipitation	7	511.2	23.1	0.0
Summer precipitation	1	511.3	23.1	0.0
Winter precipitation	7	511.3	23.1	0.0

Nitrogen

Explanatory variable	Period	AICc	ΔAICc	AICc weights
NO accumulation	-	488.1	0.0	0.6
NH accumulation	-	489.0	0.9	0.4

Linear mixed effects models (Gaussian family error distribution) analysing the relationship between average velocity of species richness changes and the change in potential explanatory variables (temperature, precipitation and nitrogen deposition). The implemented model formula was lmer($\Delta SR/\Delta t \sim \Delta T/\Delta t + \Delta P/\Delta t + N_{accum} +$ richness $+ (1|$mountain$)$). Within each new model, the focal variable (left column) was exchanged, while the remaining variables were held constant. Variables were calculated as the mean value across a period before the survey (Period). The analyses were performed with the same data as in Extended Data Table 3.

Extended Data Table 6 | Model evaluation for different time lags

Summer Temperature (15-year mean)

Time lag	AICc	ΔAICc	AICc weights
0	496.5	0.0	1.0
5	531.3	34.8	0.0
10	546.5	50.0	0.0

Annual precipitation (1-year mean)

Time lag	AICc	ΔAICc	AICc weights
0	507.7	0.0	0.72
5	510.7	3.3	0.16
10	511.3	3.6	0.12

Nitrogen accumulation

Time lag	AICc	ΔAICc	AICc weights
0	488.1	0.0	0.34
5	488.2	0.1	0.33
10	488.2	0.1	0.33

Linear mixed effects models (Gaussian family error distribution) analysing the relationship between average velocity of species richness changes and the change in potential explanatory variables (temperature, precipitation and nitrogen deposition). The implemented model formula was lmer(ΔSR/$\Delta t \sim \Delta T/\Delta t + \Delta P/\Delta t + N_{accum}$ + richness + (1|mountain)). Explanatory variables were calculated as the mean value across a period before the survey. Within each new model, the focal explanatory variable implemented with a differing time lag (time between the period and survey; left column) was exchanged while the remaining variables were held constant. The analyses were performed with the same data as in Extended Data Table 3.

Extended Data Table 7 | Trait differences between colonizing and old-established species

Trait	Fixed effect (coefficients ±std. error)		Random effect (std. deviations)
	Intercept	Difference of colonizer relative to established species	Resurvey
Plant height	-0.234 ±0.022***	+ 0.292 ±0.022***	0.31
SLA	-0.077 ±0.017***	+ 0.158 ±0.024***	0.13
Seed mass	-0.014 ±0.017	+ 0.003 ±0.025	0.09
Temperature indicator	-0.188 ±0.023***	+ 0.221 ±0.020***	0.35

Linear mixed effects models (Gaussian family error distribution) revealed systematic trait differences between colonizing and old-established species. Analyses were implemented for 364 resurveys (12,738 observations with 815 species) with a random effect of resurvey. Temperature indicator values[49] were available for 90%, specific leaf area (SLA)[48] for 61%, plant height[48] for 76%, and seed mass[48] for 53% of the observations. Significant effects are indicated by asterisks (***$P < 0.001$). Trait raw data were first log-transformed, then centred to zero mean and scaled to s.d. = 1 before analysis.

LETTER

A small peptide modulates stomatal control via abscisic acid in long–distance signalling

Fuminori Takahashi[1,2]*, Takehiro Suzuki[3], Yuriko Osakabe[1,4], Shigeyuki Betsuyaku[5,6,8], Yuki Kondo[6], Naoshi Dohmae[3], Hiroo Fukuda[6], Kazuko Yamaguchi–Shinozaki[7] & Kazuo Shinozaki[1,2]*

Mammalian peptide hormones propagate extracellular stimuli from sensing tissues to appropriate targets to achieve optimal growth maintenance[1]. In land plants, root-to-shoot signalling is important to prevent water loss by transpiration and to adapt to water-deficient conditions[2,3]. The phytohormone abscisic acid has a role in the regulation of stomatal movement to prevent water loss[4]. However, no mobile signalling molecules have yet been identified that can trigger abscisic acid accumulation in leaves. Here we show that the CLAVATA3/EMBRYO-SURROUNDING REGION-RELATED 25 (CLE25) peptide transmits water-deficiency signals through vascular tissues in *Arabidopsis*, and affects abscisic acid biosynthesis and stomatal control of transpiration in association with BARELY ANY MERISTEM (BAM) receptors in leaves. The *CLE25* gene is expressed in vascular tissues and enhanced in roots in response to dehydration stress. The root-derived CLE25 peptide moves from the roots to the leaves, where it induces stomatal closure by modulating abscisic acid accumulation and thereby enhances resistance to dehydration stress. BAM receptors are required for the CLE25 peptide-induced dehydration stress response in leaves, and the CLE25–BAM module therefore probably functions as one of the signalling molecules for long-distance signalling in the dehydration response.

The *Arabidopsis* genome contains more than 7,000 small open reading frames with no known functional annotations[5–7]. Several secreted peptides mediate cellular development in plants[8–10]. However, it is unclear whether peptide hormones mediate long-distance signalling in response to abiotic stress. CLAVATA3 (CLV3) is a well-characterized plant peptide involved in shoot apical meristem formation[11]. The *Arabidopsis* genome contains 32 CLAVATA3/EMBRYO-SURROUNDING REGION-related (CLE) family genes[12]. Tracheary element differentiation inhibitory factor (TDIF) is involved in the formation of vascular tissue that functions in water transport[13,14].

To determine whether CLE peptides modulate long-distance signalling in dehydration stress response, the induction of *NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 3* (*NCED3*), a gene that encodes a key enzyme for abscisic acid (ABA) synthesis under conditions of dehydration stress in leaves[15,16], was analysed by treating roots with 27 synthetic CLE peptides. Of these peptides, the application of CLE25 to roots induced *NCED3* expression and enhanced ABA accumulation in leaves (Extended Data Fig. 1a–c). Stomatal response to the application of CLE25, CLV3, CLE46 and TDIF peptides was analysed using CLE26 as a negative control, as because this peptide is the most homologous to CLE25 and mediates root development[17–20] (Extended Data Fig. 1d). CLE25 application induced a level of stomatal closure similar to that induced by ABA application (Fig. 1a and Extended Data Fig. 1e), whereas CLE26, CLV3, CLE46 and TDIF application did not induce stomatal closure (Extended Data Fig. 1f). Analysis of the dose-dependent effect of CLE25 on stomatal response

showed that CLE25 functions in stomatal closure at nanomolar concentrations (Fig. 1b), thereby indicating that CLE25 also functions as a hormone. Nano-liquid chromatography–tandem mass spectrometry (nLC–MS/MS) was performed using leaves from plants with roots treated with a non-labelled or isotope-labelled synthetic CLE25 peptide. Accumulation of this peptide was detected in leaves (Extended Data Fig. 1g, h), indicating that synthetic CLE25 moved from roots to leaves and acted as a functional mobile signal.

CLE25 expression was enhanced in roots in response to dehydration stress (Fig. 2a). Many genes involved in ABA biosynthesis and transport are expressed in vascular tissues[16,21]. Thus, vascular tissue is thought to be important for ABA biosynthesis under dehydration stress conditions. β-Glucuronidase (GUS) reporter-aided histochemical analysis showed that *CLE25* promoter activity occurred in lateral roots, the root tip, vascular veins of leaves and the procambium of primary roots (Fig. 2b, c). In situ hybridization analysis also detected the expression of *CLE25* mRNA in the procambium of root vascular tissues (Fig. 2d).

Using CRISPR–Cas9 genome editing, knockout mutants of the *CLE25* gene (*cle25*) were generated. A CEL-I assay of the mutant genomes confirmed a CRISPR–Cas9-mediated mutation (Extended Data Fig. 2a). Isolated *cle25* homozygous mutants contained a nonsense mutation at the guanosine at position 22 in the *CLE25* coding region (Extended Data Fig. 2b). Dehydration-induced *NCED3* expression was repressed in two *cle25* mutants (Fig. 3a). The *clv3-8*,
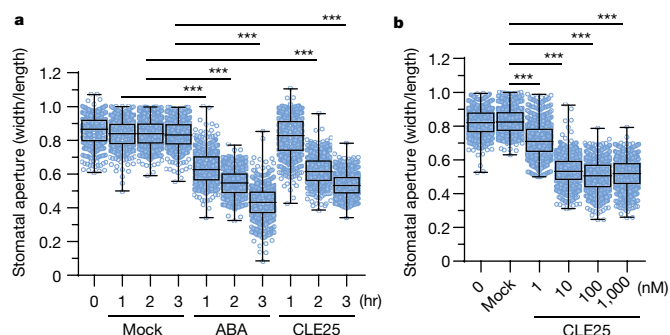
Fig. 1 | The application of the CLE25 peptide affects stomatal closure in leaves. a, Roots of whole plants were incubated without (0 h on *x* axis, *n* = 401) or with 0.01% acetonitrile (mock: *n* = 432, 1 h; *n* = 551, 2 h; *n* = 506, 3 h), ABA (*n* = 586, 1 h; *n* = 548, 2 h; *n* = 514, 3 h) or CLE25 peptide (*n* = 583, 1 h; *n* = 509, 2 h; *n* = 544, 3 h) solution for the times indicated. Data are from three experiments. **b**, Detached rosette leaves were incubated without (0 h on *x* axis, *n* = 628) or with mock (*n* = 583) or with each concentration of CLE25 peptide (*n* = 647, 1 nM; *n* = 644, 10 nM; *n* = 664, 100 nM; *n* = 622, 1,000 nM) for 3 h. Data are from three experiments. ***P < 0.001 as analysed by one-way ANOVA followed by a Tukey–Kramer post hoc test (**a**, **b**).

[1]Gene Discovery Research Group, RIKEN Center for Sustainable Resource Science, Tsukuba, Japan. [2]Biomass Research Platform Team, RIKEN Center for Sustainable Resource Science, Tsukuba, Japan. [3]Biomolecular Characterization Unit, RIKEN Center for Sustainable Resource Science, Wako, Japan. [4]Faculty of Bioscience and Bioindustry, Tokushima University, Tokushima, Japan. [5]Japan Science and Technology Agency (JST), PRESTO, Kawaguchi, Japan. [6]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan. [7]Graduate School of Agricultural and Life Science, The University of Tokyo, Tokyo, Japan. [8]Present address: Faculty of Life and Environmental Science, University of Tsukuba, Tsukuba, Japan. *e-mail: fuminori.takahashi@riken.jp; kazuo.shinozaki@riken.jp
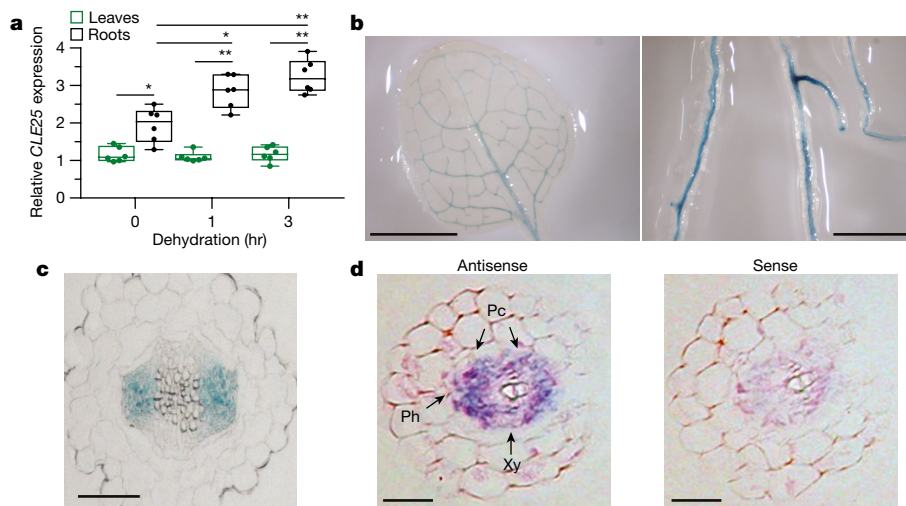
Fig. 2 | CLE25 is expressed in vascular tissues of roots and leaves.
a, Tissue-specific expression of *CLE25* in response to dehydration stress
($n = 6$ biological replicates). *$P < 0.05$, **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test. **b**, Tissue specificity of GUS staining in the roots and leaves of pCLE25::GUS plants ($n = 8$ transgenic plants). Scale bar, 0.5 mm. **c**, GUS staining of procambium in

the differentiation zone of root of a pCLE25::GUS plant ($n = 4$ transgenic plants). Scale bar, 50 μm. **d**, RNA in situ hybridization with the *CLE25* antisense or sense probe in cross sections of roots of wild-type plants ($n = 12$ biological replicates). Pc, procambium; Ph, phloem; Xy, xylem. Scale bars, 50 μm.

cle46-1 and *tdr-1* (a mutant of *TDIF RECEPTOR* (*TDIF*, also known as *PXY*)) mutants did not repress *NCED3* expression, as compared with that of wild-type plants in response to dehydration stress (Extended Data Fig. 2c–e). ABA-inducible genes—one from the LATE EMBRYOGENESIS ABUNDANT family (*AT3G02480*, here referred to as *LEA*) as well as *RESPONSIVE TO DESICCATION 29B* (*RD29B*; also known as *LTI65*)—were also suppressed in two *cle25* mutants in response to dehydration stress (Fig. 3b, c), suggesting an important role for CLE25 in the regulation of *NCED3* and ABA-induced gene expression. *NCED3* expression was strongly correlated with ABA accumulation under dehydration stress conditions. After dehydration for three hours, ABA levels increased by sixfold in the leaves of wild-type

plants (Fig. 3d), whereas ABA accumulation in roots of wild-type plants was one-tenth of that in leaves under control conditions, and did not increase in response to dehydration stress. Accumulation of ABA in response to dehydration stress was abolished in leaves of *cle25* mutants. Induction of stomatal closure by treatment with the CLE25 peptide did not occur in the ABA-deficient mutants *nced3-2* and *aba2-1*, although ABA treatment enhanced stomatal closure, which suggests that CLE25 modulates stomatal closure in response to ABA accumulation (Extended Data Fig. 3). Stomatal conductance of *cle25* mutants was similar to that of wild-type plants, under control conditions (Extended Data Fig. 4a). Water loss was greater in *cle25* mutants than in wild-type plants within one hour, which indicates that CLE25 may modulate
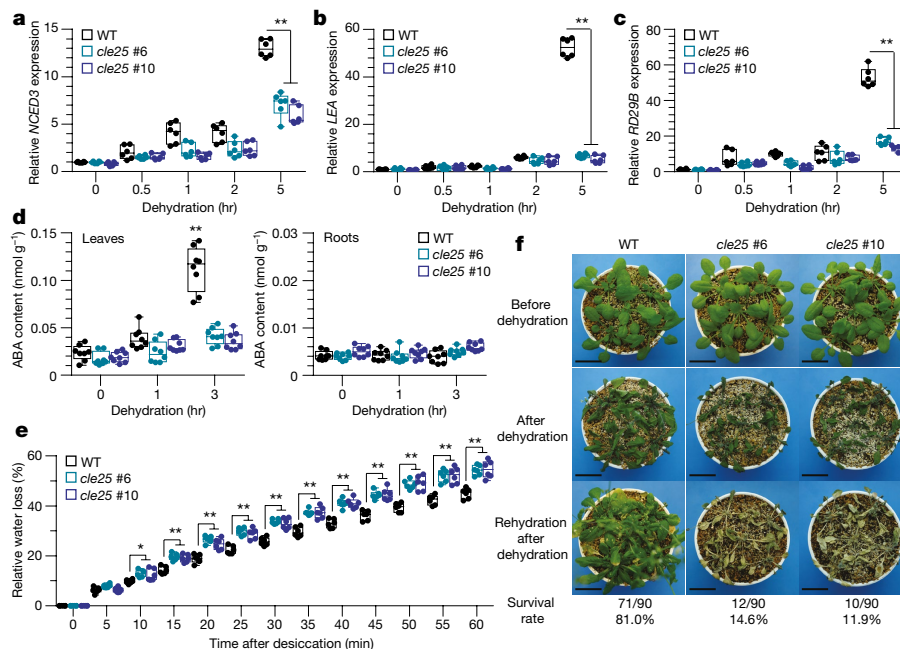
Fig. 3 | CLE25 CRISPR–Cas9-derived knockout (cle25) mutants affect expression of dehydration-induced genes, ABA accumulation, water loss and dehydration stress sensitivity. a–c, Dehydration-induced *NCED3* (**a**), *LEA* (**b**) and *RD29B* (**c**) expression in wild-type plants and *cle25* mutants ($n = 6$ biological replicate). **d**, ABA content of leaves or roots in wild-type plants and *cle25* mutants in response to dehydration stress

($n = 8$ biological replicates). **e**, Water loss in wild-type plants and *cle25* mutants in response to dehydration stress ($n = 6$ biological replicates). *$P < 0.05$, **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test (**a–e**). **f**, *cle25* mutants had a dehydration-sensitive phenotype ($n = 90$ plants per group; data from four experiments). WT, wild type. Scale bars, 2 cm.
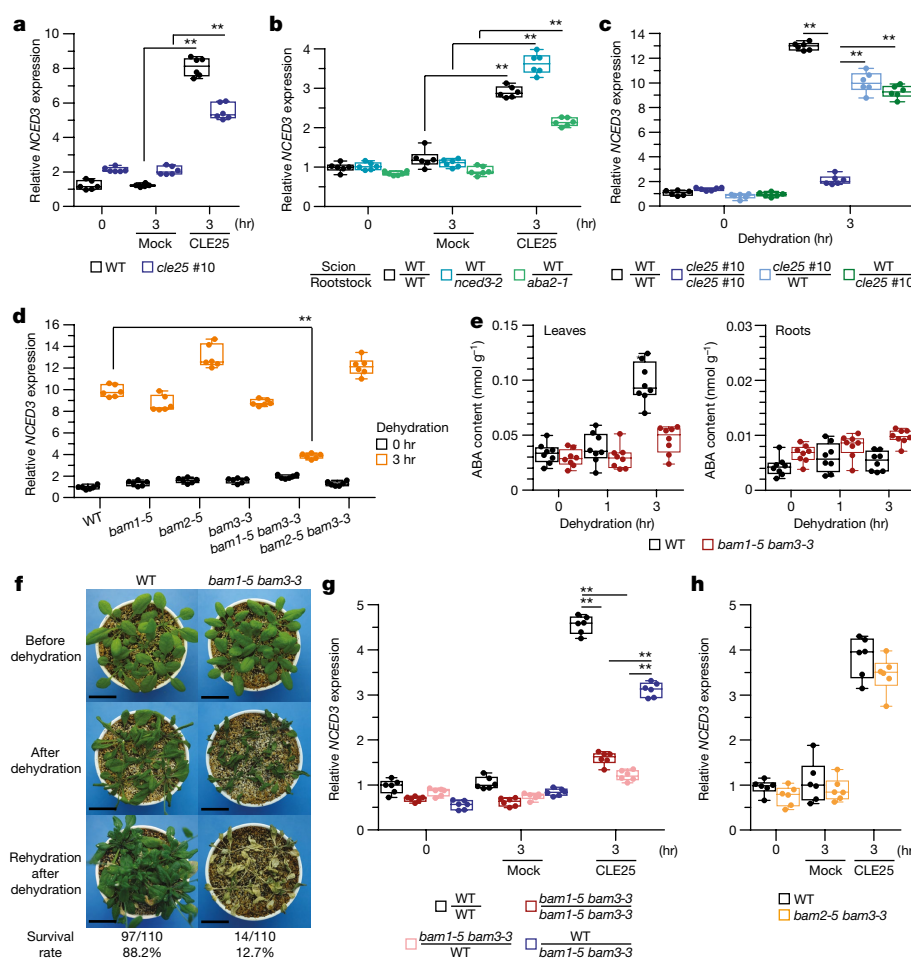
Fig. 4 | CLE25 peptide moves from roots to leaves and modulates *NCED3* expression in leaves in association with the receptor-like kinases BAM1 and BAM3. a, *NCED3* expression in the leaves of wild-type and *cle25* mutants ($n = 6$ biological replicates) after CLE25 application to roots. **b**, *NCED3* expression in grafted leaves of wild-type/wild-type, wild-type/*nced3-2* and wild-type/*aba2-1* plants ($n = 6$ biological replicates). **c**, Dehydration-induced *NCED3* expression in grafted leaves in which shoots and roots were grafted between wild-type or *cle25* #10 ($n = 6$ biological replicates). **d**, Dehydration-induced *NCED3* expression in the *bam* mutants ($n = 6$ biological replicates). **e**, ABA content of leaves and roots in wild-type plants and *bam1-5 bam3-3* mutants in response to dehydration stress ($n = 8$ biological replicates). **f**, *bam1-5 bam3-3* mutants had a dehydration-sensitive phenotype ($n = 110$ plants per group; data from four experiments). Scale bars, 2 cm. **g**, *NCED3* expression in grafted leaves, in which the shoots and roots were grafted between wild-type or *bam1-5 5 bam3-3* ($n = 6$ biological replicates). **h**, *NCED3* expression in the leaves of wild-type plants and *bam2-5 bam3-3* mutants ($n = 6$ biological replicates). **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test (**a**–**e**, **g** and **h**).

stomatal responses in association not only with ABA accumulation but also with other rapid signals, including hydraulic water tension in response to dehydration stress (Fig. 3e). The *cle25* mutants grew normally in soil under well-watered conditions (Extended Data Fig. 4b–d), but exhibited a marked sensitivity to dehydration stress (Fig. 3f). *CLE25* RNA interference (*CLE25* RNAi)-knockdown plants were also generated, in which *CLE25* expression was consistently repressed (Extended Data Fig. 5a). In *CLE25* RNAi plants, expression levels of *NCED3*, *LEA* and *RD29B* were repressed when compared with those of wild-type plants (Extended Data Fig. 5b–d). *CLE25* RNAi plants exhibited a dehydration-sensitive phenotype, which was similar to that of the *nced3-2* mutant (Extended Data Fig. 5e). These findings suggest the functional importance of CLE25 in dehydration stress responses and tolerance.

Post-translational processing is critical for the biological functions of CLE peptides. Recent studies have demonstrated that mature CLE peptides undergo post-translational modifications, such as arabinosylation or the addition of hydroxyl residues in proline[22]. Extracellular secretion of the CLE25 peptide with two hydroxyproline residues was detected in the culture medium of *Arabidopsis* T87 cells (Extended Data Fig. 6).

The application of CLE25 to roots increased *NCED3* expression in leaves of both *cle25* mutants and *CLE25* RNAi plants, indicating that CLE25 application modulates *NCED3* expression in leaves without

endogenous CLE25 peptide expression (Fig. 4a and Extended Data Fig. 7a, b). Wild-type shoot scions grafted onto *nced3-2* or *aba2-1* rootstocks exhibited an enhanced level of *NCED3* expression in leaves after CLE25 application to roots, which was similar to that of control graft plants (wild type/wild type; scion/rootstock) (Fig. 4b), suggesting that CLE25—but not root-derived ABA—modulates *NCED3* expression in leaves.

Grafted *cle25* #10/*cle25* #10 (numbers with # indicate the independent line number of *cle25* mutants) plants exhibited a reduction in *NCED3* expression in leaves in response to dehydration stress, whereas clear *NCED3* induction by dehydration was observed in wild-type/wild-type plants (Fig. 4c). Compared with *cle25* #10/*cle25* #10 plants, *cle25* #10/wild-type plants exhibited enhanced *NCED3* expression in dehydrated leaves, with expression levels reaching about 80% of those in wild-type/wild-type plants, which suggests that root-derived endogenous CLE25 modulates *NCED3* expression in leaves. Wild-type/*cle25* #10 plants exhibited dehydration-induced *NCED3* expression in a similar manner to the *cle25* #10/wild-type plants. This indicates that the *CLE25* gene in the shoot, which is not induced in response to dehydration (Fig. 2a), is also sufficient to induce an *NCED3* expression level of about 80% of that of wild-type/wild-type plants. Plants obtained by the reciprocal grafting of wild-type or *CLE25* RNAi (*CLE25* RNAi #13)

shoot scions onto wild-type or *CLE25* RNAi #13 rootstocks had similar *NCED3* expression patterns in their leaves (Extended Data Fig. 7c, d). nLC–MS/MS analyses showed the accumulation of endogenous CLE25 in dehydrated leaves of *cle25* #10/wild-type plants (Extended Data Fig. 8). These results indicate that CLE25 contributes to long-distance cellular communication.

To identify receptor-like kinases that can recognize CLE25 and modulate *NCED3* expression, several mutants of candidate receptor-like kinases related to CLE peptides—including CLV and BAM receptors—were selected. The *bam1-5 bam3-3* mutant repressed *NCED3* expression (Fig. 4d). The accumulation of ABA in response to dehydration stress was abolished in leaves of the *bam1-5 bam3-3* mutant (Fig. 4e). The *bam1-5 bam3-3* mutant exhibited a marked sensitivity to dehydration stress (Fig. 4f). Salinity stress responses partially mediate signalling in common with dehydration stress responses. The *cle25* and *bam1-5 bam3-3* mutants also exhibited sensitivity to salinity stress (Extended Data Fig. 9). CLE25 application to leaves did not increase *NCED3* expression in the leaves of *bam1-5 bam3-3* mutants (Extended Data Fig. 10a). CLE25 application to roots did not increase *NCED3* expression in leaves of grafted *bam1-5 bam3-3*/*bam1-5 bam3-3* and *bam1-5 bam3-3*/wild-type plants, although grafted wild-type/wild-type and wild-type/*bam1-5 bam3-3* plants did exhibit enhanced *NCED3* expression (Fig. 4g). Confocal sectional analysis of *cle25* and *bam1-5 bam3-3* mutants suggested that endogenous CLE25, and BAM1 and BAM3, do not affect protoxylem and metaxylem vessel formation and the vascular development of leaves and roots (Extended Data Fig. 10b–m). The application of CLE25 to roots increased *NCED3* expression in leaves of *bam2-5 bam3-3* mutants (Fig. 4h). These results suggest that the movement of CLE25 from roots to leaves modulates *NCED3* expression in association with BAM1 and BAM3 receptors in leaves. By contrast, BAM1 and BAM2 mainly mediate leaf development[23]. However, each receptor may mediate different functions under dehydration stress conditions.

These findings demonstrate that the CLE25–BAM pair functions as one of the modules of long-distance signalling in response to dehydration stress. The application of CLE25 and CLE26 inhibits root growth but not protoxylem vessel formation, suggesting that CLE25- and CLE26-induced root growth inhibition is different from that of other CLE peptides[13,20,24]. This root growth inhibition was an ABA-independent effect (Extended Data Fig. 10n–s). By contrast, the application of CLE25 modulates dehydration responses in leaves, but the application of CLE26 does not (Fig. 1 and Extended Data Fig. 1a–f). Physiological functions of CLE25 may be controlled in different target tissues. *Lotus japonicus* CLE-root signal proteins propagate nitrate status signals from roots to shoots[25]. Thus, peptide hormones can strongly coordinate information from the underground and aerial parts of plants. The CLE25–BAM module may transmit dehydration stress signals more precisely to specific tissues than does the ABA regulatory system.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0009-2.

1. Michell, A. R., Debnam, E. S. & Unwin, R. J. Regulation of renal function by the gastrointestinal tract: potential role of gut-derived peptides and hormones. *Annu. Rev. Physiol.* **70**, 379–403 (2008).
2. Steudle, E. The cohesion–tension mechanism and the acquisition of water by plant roots. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 847–875 (2001).
3. Christmann, A., Grill, E. & Huang, J. Hydraulic signals in long-distance signaling. *Curr. Opin. Plant Biol.* **16**, 293–300 (2013).
4. Kim, T. H., Böhmer, M., Hu, H., Nishimura, N. & Schroeder, J. I. Guard cell signal transduction network: advances in understanding abscisic acid, CO_2, and Ca^{2+} signaling. *Annu. Rev. Plant Biol.* **61**, 561–591 (2010).
5. Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H. & Shiu, S. H. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.* **17**, 632–640 (2007).
6. Hanada, K. et al. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**, 399–400 (2010).
7. Hanada, K. et al. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl Acad. Sci. USA* **110**, 2395–2400 (2013).
8. Murphy, E., Smith, S. & De Smet, I. Small signaling peptides in *Arabidopsis* development: how cells communicate over a short distance. *Plant Cell* **24**, 3198–3217 (2012).
9. Czyzewicz, N., Yue, K., Beeckman, T. & De Smet, I. Message in a bottle: small signalling peptide outputs during growth and development. *J. Exp. Bot.* **64**, 5281–5296 (2013).
10. Endo, S., Betsuyaku, S. & Fukuda, H. Endogenous peptide ligand–receptor systems for diverse signaling networks in plants. *Curr. Opin. Plant Biol.* **21**, 140–146 (2014).
11. Shinohara, H. & Matsubayashi, Y. Arabinosylated glycopeptide hormones: new insights into CLAVATA3 structure. *Curr. Opin. Plant Biol.* **13**, 515–519 (2010).
12. Betsuyaku, S., Sawa, S. & Yamada, M. The function of the CLE peptides in plant development and plant–microbe interactions. *Arabidopsis Book* 9, (e0149 (2011).
13. Ito, Y. et al. Dodeca-CLE peptides as suppressors of plant stem cell differentiation. *Science* **313**, 842–845 (2006).
14. Hirakawa, Y. et al. Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. *Proc. Natl Acad. Sci. USA* **105**, 15208–15213 (2008).
15. Iuchi, S. et al. Regulation of drought tolerance by gene manipulation of 9-*cis*-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in *Arabidopsis*. *Plant J.* **27**, 325–333 (2001).
16. Endo, A. et al. Drought induction of *Arabidopsis* 9-*cis*-epoxycarotenoid dioxygenase occurs in vascular parenchyma cells. *Plant Physiol.* **147**, 1984–1993 (2008).
17. Rodriguez-Villalon, A. et al. Molecular genetic framework for protophloem formation. *Proc. Natl Acad. Sci. USA* **111**, 11551–11556 (2014).
18. Rodriguez-Villalon, A., Gujas, B., van Wijk, R., Munnik, T. & Hardtke, C. S. Primary root protophloem differentiation requires balanced phosphatidylinositol-4,5-biphosphate levels and systemically affects root branching. *Development* **142**, 1437–1446 (2015).
19. Czyzewicz, N. et al. Modulation of *Arabidopsis* and monocot root architecture by CLAVATA3/EMBRYO SURROUNDING REGION 26 peptide. *J. Exp. Bot.* **66**, 5229–5243 (2015).
20. Kinoshita, A. et al. Gain-of-function phenotypes of chemically synthetic CLAVATA3/ESR-related (CLE) peptides in *Arabidopsis thaliana* and *Oryza sativa*. *Plant Cell Physiol.* **48**, 1821–1825 (2007).
21. Kuromori, T. et al. ABC transporter AtABCG25 is involved in abscisic acid transport and responses. *Proc. Natl Acad. Sci. USA* **107**, 2361–2366 (2010).
22. Matsubayashi, Y. Posttranslationally modified small-peptide signals in plants. *Annu. Rev. Plant Biol.* **65**, 385–413 (2014).
23. DeYoung, B. J. et al. The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in *Arabidopsis*. *Plant J.* **45**, 1–16 (2006).
24. Kondo, Y., Hirakawa, Y., Kieber, J. J. & Fukuda, H. CLE peptides can negatively regulate protoxylem vessel formation via cytokinin signaling. *Plant Cell Physiol.* **52**, 37–48 (2011).
25. Okamoto, S., Shinohara, H., Mori, T., Matsubayashi, Y. & Kawaguchi, M. Root-derived CLE glycopeptides control nodulation by direct binding to HAR1 receptor kinase. *Nat. Commun.* **4**, 2191 (2013).

Author contributions F.T. and K.S. designed the study. S.B., H.F. and F.T. prepared synthetic peptides and performed peptide screening. F.T., S.B. and K.S. performed stomatal aperture analyses. F.T., T.S. and N.D. conducted nLC–MS/MS analyses. S.B. helped with peptide labelling experiments. Y.O. and F.T. generated knockout mutants. F.T. and Y.O. performed drought stress sensitivity and stomatal conductance analyses. F.T. performed gene expression analyses and ABA measurement. Y.K., H.F. and F.T. performed structural analyses of vasculature. F.T., T.S., Y.O., S.B. and N.D. analysed the data. F.T., K.Y.-S. and K.S. wrote the manuscript with substantial input from S.B. and H.F. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

METHODS

Plant materials and growth conditions. *Arabidopsis thaliana* ecotype Col-0 was used as a wild-type control and as the genetic background of the transgenic lines. Wild-type, transgenic and mutant plants were grown on germination medium agar plates containing 1% sucrose under long-day conditions (16 h light:8 h dark) at 22 °C. *nced3-2* was provided by K. Urano (RIKEN Center for Sustainable Resource Science). The *cle46-1* (SALK_207109C), *bam1-5* (SALK_152555), *bam2-5* (GK-791G02) and *bam3-3* (SALK_118860) mutant seeds were obtained from the *Arabidopsis* Biological Resource Center (Columbus). *A. thaliana* T87 cells were cultured in 100 ml of Jouanneau and Péaud-Lenoël (JPL) medium[26] with gentle agitation (110 r.p.m.) under continuous illumination at 22 °C. A 1-ml aliquot of the cell suspension was transferred to fresh medium every 10 days.

Synthetic CLE peptide treatment. All synthesized CLE peptides contained two hydroxyproline residues[13]. The purity of each synthesized peptide was > 95%. At 25 days after germination, seedlings of control and grafted plants were transferred to 3 ml of water for 16 h. Water was then replaced with opening buffer (20 mM KCl, 1 mM CaCl$_2$ and 5 mM 2-(N-morpholino)ethanesulfonic acid-KOH, pH 6.15) and incubated for 2 h under light conditions (180–200 μmol photons s^{-1} m^{-2}). Plant roots (as shown in Fig. 4a, b, g, h and Extended Data Figs. 1b, c, 7a, b) or detached leaves (as shown in Extended Data Fig. 10a) were then transferred to sterile water containing 1 μM synthetic CLE peptides, 0.01% acetonitrile solution as mock treatment, 5 μM synthetic CLE peptides or 0.05% acetonitrile solution as mock treatment (as shown in Extended Data Fig. 1a) and incubated for different times as indicated in the figures. All rosette leaves were collected for gene expression analysis by quantitative RT–PCR (as shown in Fig. 4a, b, g, h and Extended Data Figs. 1a, b, 7a, b, 10a).

Measurement of stomatal apertures. Four-week-old soil-grown *Arabidopsis* plants or rosette leaves detached from three-week-old soil-grown *Arabidopsis* plants were placed on glass slides in opening buffer with abaxial sides facing up for 2 h under light conditions (180–200 μmol photons s^{-1} m^{-2}) to open the stomata. Roots of whole plants or detached leaves were then transferred to sterile water containing 1 μM ABA, 1 μM synthetic peptides or 0.01% acetonitrile solution as mock, and then incubated for the times indicated, to close the stomata. To analyse intact plant leaves, images of stomatal apertures were obtained using Suzuki's Universal Micro-Printing (SUMP) method with SUMP liquid and SUMP plate C (SUMP Laboratory). SUMP images were observed using an AxioPlan 2 Microscope System (Carl Zeiss). Width and length of stomatal aperture were measured with PhotoRuler version 1.1.3. Stomatal opening was calculated by the width:length ratio.

Purification of endogenous CLE25 peptides. On day 7 after subculture, cultured *Arabidopsis* T87 cells were subjected to stress treatment. Cell concentration was adjusted to 150–200 mg ml^{-1}. Stress was applied by adding JPL medium with or without 1 M mannitol for 4 h. The final mannitol concentration was 0.4 M. After treatment, the liquid culture medium was filtered through a polyethersulfone membrane (0.1-μm pore size). Peptides were purified from the culture medium by ion-exchange and reversed-phase chromatography, followed by solid-phase extraction. The dried medium was dissolved in 300 ml of water and subjected to solid-phase extraction (InertSep C18-B 60 mL; GL Sciences). After washing with aqueous 0.1% trifluoroacetic acid, the sample was eluted with 15% acetonitrile from the same solution. The eluate was concentrated by lyophilization and applied to an ion-exchange column (PolySULFOETHYL-A, 2.1 × 100 mm; Poly-LC), equilibrated with 20 mM sodium phosphate buffer (pH 3.0) at a flow rate of 0.1 ml min^{-1} and eluted with a linear gradient of 3 ml of 0–0.5 M NaCl as the equilibration buffer. The fraction eluted at the retention time corresponding to the synthetic CLE25 peptide was collected, then purified further by reversed-phase chromatography with an Inertsil ODS-3 column (1 × 100 mm; GL Sciences) and eluted with a 12.5–50% gradient of acetonitrile in aqueous 0.1% trifluoroacetic acid over a period of 60 min.

Arabidopsis leaves that absorbed 5 μM non-labelled or isotope-labelled CLE25 peptide from roots for 3 h, or leaves of wild-type or *cle25* #10 shoot scions grafted onto wild-type rootstocks under 3 h of dehydration stress were used for peptide extraction. Peptides were fractionated with an Amicon Ultra 10 K filter (Merck Millipore) and then purified using GL-Tip SDB and GC columns (GL Sciences).

Mass spectrometry analysis. To analyse the endogenous CLE25 peptide from *Arabidopsis* T87 suspension-culture cells, the fraction collected at the retention time corresponding to the synthetic CLE25 peptide was subjected to nLC–MS/MS using a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Peptides were separated using a nano-ESI spray column (100-mm length × 75-μm internal diameter, 3-μm opening, NTCC analytical column C18; Nikkyo Technos), which was equilibrated with buffer A (0.1% aqueous formic acid) and eluted with a linear gradient of 30% buffer B (0.1% formic acid in 100% acetonitrile) over a 20-min period at a flow rate of 300 nl min^{-1} (Easy nLC; Thermo Fisher Scientific). The mass spectrometer was operated in positive-ion mode and MS/MS spectra were acquired in targeted MS/MS mode ($m/z = 459.24$ and 688.36).

To detect CLE25 peptide, nLC–MS/MS analyses were conducted with the following instruments: an Autosampler-2 1D plus, NanoLC Ultra (Eksigent Technologies) and TripleTOF 5600 (SCIEX). An L-column ODS C-18 (5-mm length × 0.3-mm internal diameter, 5-μm opening) was used as a sample trap and an L-column Micro C-18 (150-mm length × 75-μm internal diameter, 5-μm opening) was used to prepare peptide samples (Chemical Evaluation and Research Institute). The injection volume was 1 μl and the flow rate was 300 nl min^{-1}. The mobile phases comprised 2% acetonitrile and 0.1% formic acid (A) and 80% acetonitrile and 0.1% formic acid (B). The linear gradient comprised A:B = 98:2 to A:B = 60:40 for 125 min, A:B = 10:90 for 5 min and A:B = 98:2 for 20 min. An ion spray voltage of 1400 V was applied via the metal connector with a Dream Spray closed-type nanospray source (AMR). The MS scan ranged from an m/z of 400–1250. The top 10 precursor ions were selected for subsequent MS/MS scans in the high-sensitivity mode. Peaks at m/z 688.34 $[\text{M} + 2\text{H}]^{2+}$ (non-labelled CLE25 peptide), 691.35 $[\text{M} + 2\text{H}]^{2+}$ (isotope-labelled CLE25 peptide) or 459.24 $[\text{M} + 3\text{H}]^{3+}$ (endogenous CLE25 peptide), and the product ions were detected by nLC–MS/MS.

Extended Data Fig. 6c lists the top 10 proteins with the highest accumulation in T87 cells in response to mannitol treatment. Concentrations of these proteins in T87 cells in response to mannitol conditions indicate similar accumulations in T87 cells under control conditions. Columns of the accumulation in liquid culture medium indicate the amounts of these proteins in the liquid culture medium of T87 cells cultured with or without mannitol. Accumulations of these proteins in the liquid culture medium were similar under control conditions and in response to mannitol treatment.

Plasmid construction. The vector pGreenII 0129, which contained the CaMV35S promoter and the PDK intron of pKANNIBAL, was used to prepare the RNAi construct. A 215-bp fragment (5′-ATGCTTGTTTTTTGCTTCCCATTT CGCTTTCCCCTTTTTTGAGCCTCTTCTGTCCAAAGATATCTCTCTCTA TTTATGTGACAGTCACTTCACCAACATCATGGATGTTCTGCTCAGTTT ATTCTTGGGTTTGGTTGGTCAGTTGTTTATGTTAAACAGGAAGCTGTAG GGACATAGGTTTCAGTATGGGTGGAAATGGCATTAGAGCTTTGGTT-3′) corresponding to the leader sequence and first exon of *CLE25* was isolated via PCR with the incorporated restriction sites, which produced compatible ends. The CasOT algorithm was used to design suitable guide RNAs (gRNAs) without off-targets via the website 'focas' (http://focas.ayanel.com/)[27,28]. The designed 18-bp gRNA (5′-GAAATGGCATTAGAGCTT-3′) was inserted into the *Bsa*I restriction site in the CRISPR–Cas9 vector, pEgP526-2A-GFBSD2 plasmid or pEgP126_Paef1-2A-GFPSD2 plasmid[28]. These plasmids were transformed into *Agrobacterium tumefaciens* strain GV3101 and then introduced into *Arabidopsis* using the floral dipping method.

Analysis of mutations at CRISPR–Cas9 target sites. Genomic DNA was extracted from each of the CRISPR–Cas9 transgenic plants and wild-type plants. A CEL-1 assay was performed with a Guide-it Mutation Detection Kit (Takara Bio) using the 300-bp PCR products surrounding the target locus of the gRNA from the selected T1 plants containing pEgP526-2A-GFBSD2 or wild-type plants, with PrimeSTAR GXL DNA polymerase (Takara Bio). After confirming the presence of the mutation induced by pEgP526-2A-GFBSD2, the gRNA was introduced into pEgP126_Paef1-2A-GFPSD2 to obtain an early generation of bi-allelic mutants[28]. Sequences of PCR products and sub-clones obtained using DNA from T2 plants were analysed to determine segregation of mutants in T2 and T3 plants, and T3 homozygous mutants were isolated for further analysis.

Quantitative RT–PCR analysis. Total RNA was isolated using a Trizol-modified reagent. First-strand cDNA was synthesized from 5 μg of total RNA using random hexamer primers and SuperScript III reverse transcriptase (Invitrogen Corporation). Quantitative RT–PCR was performed with gene-specific primers according to instructions provided with Primer Express Software version 3.0.1 (Life Technologies Corporation) and SYBR Premix Ex Taq (Takara Bio) and analysed using a 7500 Fast Real-Time PCR system (Applied Biosystems) with the following gene-specific primer sets: *CLE25* forward, 5′-GGTAAGGATGTGAATCTGTTTCATGT-3′; *CLE25* reverse, 5′-TCTGCTTTCCTGTTGTGGATAGG-3′; *NCED3* forward, 5′-CACGATTTCGCGATTACAGAGA-3′; *NCED3* reverse, 5′-CCGGCAGCTTGAAAACGA-3′; *LEA* forward, 5′-GCAAAA CGCGAGCTACCAA-3′; *LEA* reverse, 5′-GTCCAGTCTGTTGC AAGGAGTCT-3′; *RD29B* forward, 5′-GCGCACCAGTGTATGAATCCT-3′; *RD29B* reverse, 5′-CGGCATGACTAAGAGACTTAGGTTT-3′; and *Actin2* forward, 5′-AGTGGTCGTACAACCGGTATTGT-3′; *Actin2* reverse, 5′-GATGGCATGAGGAAGAGAGAAAC-3′.

Histochemical analysis of GUS expression in transgenic plants. Transgenic *Arabidopsis* plants harbouring the GUS reporter gene fused to the 1.19-kb *Arabidopsis CLE25* promoter were used for histochemical GUS assays. To analyse *CLE25* expression, 4-, 9-, 12- or 16-day-old seedlings were stained with a GUS staining buffer (100 mM Tris-HCl, pH 7.0, 2 mM ferricyanide and 1 mM 5-bromo-4-chloro-3-indolyl-β-d-glucuronidase) for 16 h at 37 °C. GUS expression

profiles were then determined using a microscope. The images represent 9-day-old seedlings. The roots of transgenic plants were fixed in a 1:3 mixture of acetic acid:ethanol and mounted in a mixture of chloral hydrate:glycerol:water (8:1:2). The fixed roots were embedded in Technovit 7100 (Heraeus Kulzer) after the series of ethanol replacement for generating cross sections. Sections were made at 5-μm thickness by Microtome (HM335E; MICROM GmbH).

In situ hybridization analysis. *Arabidopsis* tissues were fixed with 10% neutral-buffered formalin (NBF) + 50% ethanol solution, embedded in paraffin on CT-Pro20 using G-Nox as a less-toxic organic solvent than xylene and sectioned at a thickness of 5 μm. In situ hybridization was performed with an ISH Reagent Kit (GenoStaff) according to the manufacturer's instructions. Tissue sections were deparaffinized with G-Nox and rehydrated using an ethanol series and phosphate-buffered saline (PBS). The sections were fixed with 10% formalin in PBS for 15 min at room temperature, washed in PBS, treated with 3 μg ml^{-1} of proteinase K (Wako Pure Chemical Industries) in PBS for 10 min at 37 °C, washed in PBS, re-fixed with 10% NBF for 15 min at room temperature, washed in PBS, placed in 0.2 N HCl for 10 min at room temperature, washed in PBS and placed within a Coplin jar containing 1 × G-Wash (GenoStaff), which was equal to 1 × saline sodium citrate. Hybridization was performed with probes at concentrations of 300 ng ml^{-1} in G-Hybo-L (GenoStaff) for 16 h at 60 °C and then with 50% formamide in 1 × G-Wash for 10 min at 60 °C. The sections were washed twice in 1 × G-Wash for 10 min at 60 °C, twice in 0.1 × G-Wash for 10 min at 60 °C and finally twice in Tris-buffered saline with 0.1% Tween 20 (TBST) at room temperature. After treatment with 1 × G-Block (GenoStaff) for 15 min at room temperature, the sections were incubated with an anti-DIG AP conjugate (Roche Diagnostics K.K.) diluted to 1:2000 with 50 × G-Block (GenoStaff) in TBST for 1 h at room temperature. The sections were washed twice in TBST and incubated in 100 mM NaCl, 50 mM MgCl$_2$, 0.1% Tween 20 and 100 mM Tris-HCl (pH 9.5). Colouring reactions were performed overnight with an NBT/BCIP solution (Sigma-Aldrich) and then washed with PBS. The sections were counterstained with Kernechtrot stain solution (Muto Pure Chemicals) and mounted with G-Mount (GenoStaff).

Dehydration stress treatments. To induce dehydration stress, 18-day-old seedlings of each transgenic plant or 25-day-old seedling of each grafted plant were transferred into 3 ml of water for 16 h. Plants were removed from water and then incubated for times indicated under long-day conditions (16 h light:8 h dark) at 22–25 °C in 45–60 relative humidity to induce dehydration stress. A minimum of four seedlings were used for each experimental condition. All rosette leaves (Figs. 2a, 4c and Extended Data Fig. 1a, 7d), all roots (Fig. 2a and Extended Data Fig. 7c) and all whole seedlings (Figs. 3a–c, 4d and Extended Data Fig. 2c–e, 5a–d) were collected for use in other experiments.

The amount of water loss of wild-type and mutant plants was calculated by weighing each plant at the times indicated. All changes in fresh weight are presented as percentages in Fig. 3e.

To analyse dehydration stress-sensitive phenotypes, wild-type, transgenic and mutant plants were grown in soil at 22 °C for 4 weeks under long-day conditions (illumination at 40–60 μmol photons s^{-1} m^{-2}). Water content ratio of each pot was adjusted to 59.1% (Daio Kasei, professional grove soil, 45 g; water, 65 g), and then the water supply was stopped. Each pot was rotated every half a day while water was withheld. After 14–16 days of withholding water, the water supply was restarted. Three images of the same plants before dehydration, after dehydration and rehydration after dehydration are shown. Survival rates of each genotype were measured during rehydration after dehydration.

Measurement of ABA levels. To quantify ABA levels, dehydration stress was induced in detached roots or whole seedlings by incubation for the times indicated in the figures. All rosette leaves were collected from whole seedlings to measure ABA content in leaves only. All samples were ground in liquid nitrogen. ABA was extracted with 80% methanol, 500 mg l^{-1} of citric acid and 10 mg l^{-1} of butylated hydroxytoluene. After centrifugation to remove any debris, the supernatant was dried and reconstituted with Tris-buffered saline (25 mM Tris, 100 mM NaCl, 1 mM MgCl$_2$, pH 7.5). ABA was measured with a Phytodetek ABA measurement kit (Agdia) according to the manufacturer's instructions.

Micrografting. Wild-type, transgenic and mutant plant seeds were sown in half-strength MS medium containing 0.7% agar and 0% sucrose with cellulose nitrate membranes under long-day conditions (10–20 μmol photons s^{-1} m^{-2}) at 22 °C. Hypocotyls were cut from 5-day-old seedlings of each genotype using a syringe needle. Shoot scions from wild-type, transgenic or mutant plants were reciprocally grafted onto rootstocks of wild-type, transgenic or mutant plants. Grafted plants were incubated in a half-strength MS medium containing 1.5% agar and 0% sucrose under long-day conditions (10–20 μmol photons s^{-1} m^{-2}) at 26 °C for 4 days and then incubated at 22 °C for 2 days. After growing in half-strength MS medium with 0.8% agar and 1% sucrose for 4 days, plants were transferred to charcoboll soil (IMPACK) and grown for 10 days under long-day conditions (illumination at 40–60 μmol photons s^{-1} m^{-2}).

Chlorophyll measurement. Seeds of wild-type, *cle25* mutant or *bam1-5 bam3-3* mutant plants were germinated and grown on germination medium agar plates containing different concentrations of NaCl (0, 130, 140 or 150 mM) for 16 days. Seedlings were collected and ground in liquid nitrogen. Chlorophyll was extracted with 80% acetone. Absorbance was measured at 646.6 and 663.6 nm using an EnSpire multimode plate reader (PerkinElmer, Waltham, MA, USA). Chlorophyll content was determined as: chlorophyll a + chlorophyll $b = 17.76 \times A_{646.6} + 7.34 \times A_{663.6}$. The chlorophyll content of wild-type plants grown under 0-mM-NaCl conditions was used to normalize the chlorophyll content of wild-type plants, *cle25* mutants or *bam1-5 bam3-3* mutants grown with or without NaCl.

Gas exchange measurements. Stomatal conductance was assayed in seedlings of 25-day-old wild-type plants and *cle25* mutants using a portable gas exchange system (LI-6400; LI-COR). The illumination was set at 60 μmol photons s^{-1} m^{-2}, the air flow was set to 500 μmol s^{-1} and the CO$_2$ concentration of the air was controlled at 400 p.p.m. using a CO$_2$ cylinder during experiments.

Observation of vasculature. Cotyledons and roots were fixed in a 1:3 mixture of acetic acid:ethanol and mounted in a mixture of chloral hydrate:glycerol:water (8:1:2). Vascular images were obtained with a light microscope (BX51; Olympus). For generating cross sections, fixed roots were embedded in Technovit 7100 (Heraeus Kulzer) after the series of ethanol replacement. Sections were made at 3-μm thickness by Ultramicrotome (RM2165; Leica). The sections were stained with 0.01% toluidine blue O and observed under a light microscope (BX51; Olympus).

Statistical analyses and reproducibility. All statistical tests and *n* numbers, including sample sizes or biological replications, are described in the figure legends. Central lines indicate median and variation indicates interquartile range in box-and-whisker plots. Central lines indicate median in dot plots. For comparison between two groups, two-tailed Student's *t*-test was used. For comparison between more than two groups, one-way ANOVA was used, and followed by a Tukey's or a Tukey–Kramer post hoc test. For assessment between two independent variables, two-way ANOVA was used, followed by a Tukey's post hoc test. We could not provide the exact *P* values because a one-way ANOVA was used, followed by a Tukey's or Tukey–Kramer post hoc test for comparison between more than two groups. The *t*-test analyses were performed with an alpha level of 0.01 or 0.05, and provided with the following *t*-values (*t*) and degrees of freedom (d.f.). For Extended Data Fig. 1a, dehydration, $t = 17.36$, d.f. = 6, $P = 0.00000021$; CLE25, $t = 0.38$, d.f. = 6, $P = 0.72$; CLV3, $t = 1.75$, d.f. = 6, $P = 0.13$; CLE46, $t = 1.64$, d.f. = 6, $P = 0.15$; and TDIF, $t = 1.13$, d.f. = 6, $P = 0.30$. For Extended Data Fig. 7d, wild-type/*CLE25* RNAi #13 2 h, $t = 1.17$, d.f. = 10, $P = 0.27$; wild-type/*CLE25* RNAi #13 5 h, $t = 1.21$, d.f. = 10, $P = 0.25$; and *CLE25* RNAi #13/wild-type 5 h, $t = 0.21$, d.f. = 10, $P = 0.84$. For Extended Data Fig. 9, *cle25* #10 130 mM, $t = 3.84$, d.f. = 4, $P = 0.018$; *bam1-5 bam3-3* 130 mM, $t = 4.02$, d.f. = 4, $P = 0.016$; *cle25* #10 140 mM, $t = 3.88$, d.f. = 4, $P = 0.018$; *bam1-5 bam3-3* 140 mM, $t = 5.70$, d.f. = 4, $P = 0.0026$; *cle25* #6 150 mM, $t = 4.74$, d.f. = 4, $P = 0.0090$; *cle25* #10 150 mM, $t = 8.02$, d.f. = 4, $P = 0.0013$; and *bam1-5 bam3-3* 150 mM, $t = 10.92$, d.f. = 4, $P = 0.0004$.

ANOVA analyses were performed with an alpha level of 0.01 or 0.05, and provided with the following *F* values (*F*) and degrees of freedom (d.f.). For Fig. 1a, $F = 1,605.09$, d.f. = 9. For Fig. 1b, $F = 1,845.46$, d.f. = 5. For Fig. 2a, $F = 45.35$, d.f. = 5. For Fig. 3a, $F = 114.24$, d.f. = 14. For Fig. 3b, $F = 559.38$, d.f. = 14. For Fig. 3c, $F = 148.82$, d.f. = 14. For Fig. 3d: leaves, $F = 53.141$, d.f. = 8; roots, $F = 3.29$, d.f. = 8. For Fig. 3e: 5 min, $F = 5.16$, d.f. = 2; 10 min, $F = 6.41$, d.f. = 2; 15 min, $F = 23.58$, d.f. = 2; 20 min, $F = 29.14$, d.f. = 2; 25 min, $F = 32.11$, d.f. = 2; 30 min, $F = 41.65$, d.f. = 2; 35 min, $F = 33.53$, d.f. = 2; 40 min, $F = 38.12$, d.f. = 2; 45 min, $F = 32.37$, d.f. = 2; 50 min, $F = 35.26$, d.f. = 2; 55 min, $F = 27.96$, d.f. = 2, 60 min, $F = 26.48$, d.f. = 2. For Fig. 4a, $F = 418.60$, d.f. = 5. For Fig. 4b, $F = 283.88$, d.f. = 8. For Fig. 4c, $F = 913.17$, d.f. = 7. For Fig. 4d, $F = 513.85$, d.f. = 11. For Fig. 4e: leaves, $F = 34.91$, d.f. = 5; roots, $F = 8.38$, d.f. = 5. For Fig. 4g, $F = 576.44$, d.f. = 11. For Fig. 4h, $F = 99.98$, d.f. = 5. For Extended Data Fig. 1b, $F = 50.17$, d.f. = 5. For Extended Data Fig. 1c, $F = 23.98$, d.f. = 5. For Extended Data Fig. 1f, $F = 1,713.73$, d.f. = 6. For Extended Data Fig. 2c, $F = 190.26$, d.f. = 9. For Extended Data Fig. 2d, $F = 529.81$, d.f. = 9. For Extended Data Fig. 2e, $F = 61.53$, d.f. = 9. For Extended Data Fig. 3, $F = 3,302.34$, d.f. = 8. For Extended Data Fig. 4a, $F = 17.78$, d.f. = 19. For Extended Data Fig. 4b, $F = 2.52$, d.f. = 2. For Extended Data Fig. 4c, $F = 0.092$, d.f. = 2. For Extended Data Fig. 4d, $F = 0.69$, d.f. = 2. For Extended Data Fig. 5a, $F = 159.01$, d.f. = 14. For Extended Data Fig. 5b, $F = 1,430.26$, d.f. = 14. For Extended Data Fig. 5c, $F = 1,829.29$, d.f. = 14. For Extended Data Fig. 5d, $F = 1,405$, d.f. = 14. For Extended Data Fig. 7a, $F = 107.24$, d.f. = 5. For Extended Data Fig. 7b, $F = 1,959.43$, d.f. = 5. For Extended Data Fig. 7c, $F = 70.70$, d.f. = 19. For Extended Data Fig. 10a, $F = 52.66$, d.f. = 5. For Extended Data Fig. 10n, genotypes, $F = 3.29$, d.f. = 2. For Extended Data Fig. 10p, genotypes, $F = 0.35$, d.f. = 2. For Extended Data Fig. 10r, genotypes, $F = 5.89$, d.f. = 2.

Statistical methods were used to predetermine sample size. For the samples extracted from the infinite population, such as in the analysis of stomatal aperture, we calculated and determined the sample size with Excel and Visual Basic

for Applications. The parametric test is used for other statistical analyses because the data indicate that the population is normally distributed and the population variance is equal. It is generally thought that the sample size is guaranteed with $n \geq 6$ in those tests. For these reasons, we determined the sample size of key data as $n \geq 6$ in our analysis. All samples were allocated randomly into experimental groups, and all experiments were blinded during data acquisition and analyses.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Source data for Figs. 1–4 and Extended Data Figs. 1–5, 7, 9, 10 are provided with the paper. The raw image of electrophorese are provided in Supplementary Fig. 1. Sequence data used in this paper can be found in The *Arabidopsis* Information Resource (TAIR) database (https://www.arabidopsis.org/) under the following accessions: At3g28455 (CLE25), At3g14440 (NCED3), At3g02480 (LEA), At5g52300 (RD29B) and At3g18780 (ACT2).Other data that support the findings of this study are available from the corresponding authors upon request.

26. Axelos, M., Curie, C., Mazzolini, L., Bardet, C. & Lescure, B. A protocol for transient gene expression in *Arabidopsis thaliana* protoplasts isolated from cell suspension cultures. *Plant Physiol. Biochem.* **30**, 123–128 (1992).
27. Xiao, A. et al. CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* **30**, 1180–1182 (2014).
28. Osakabe, Y. et al. Optimization of CRISPR/Cas9 genome editing to modify abiotic stress responses in plants. *Sci. Rep.* **6**, 26685 (2016).
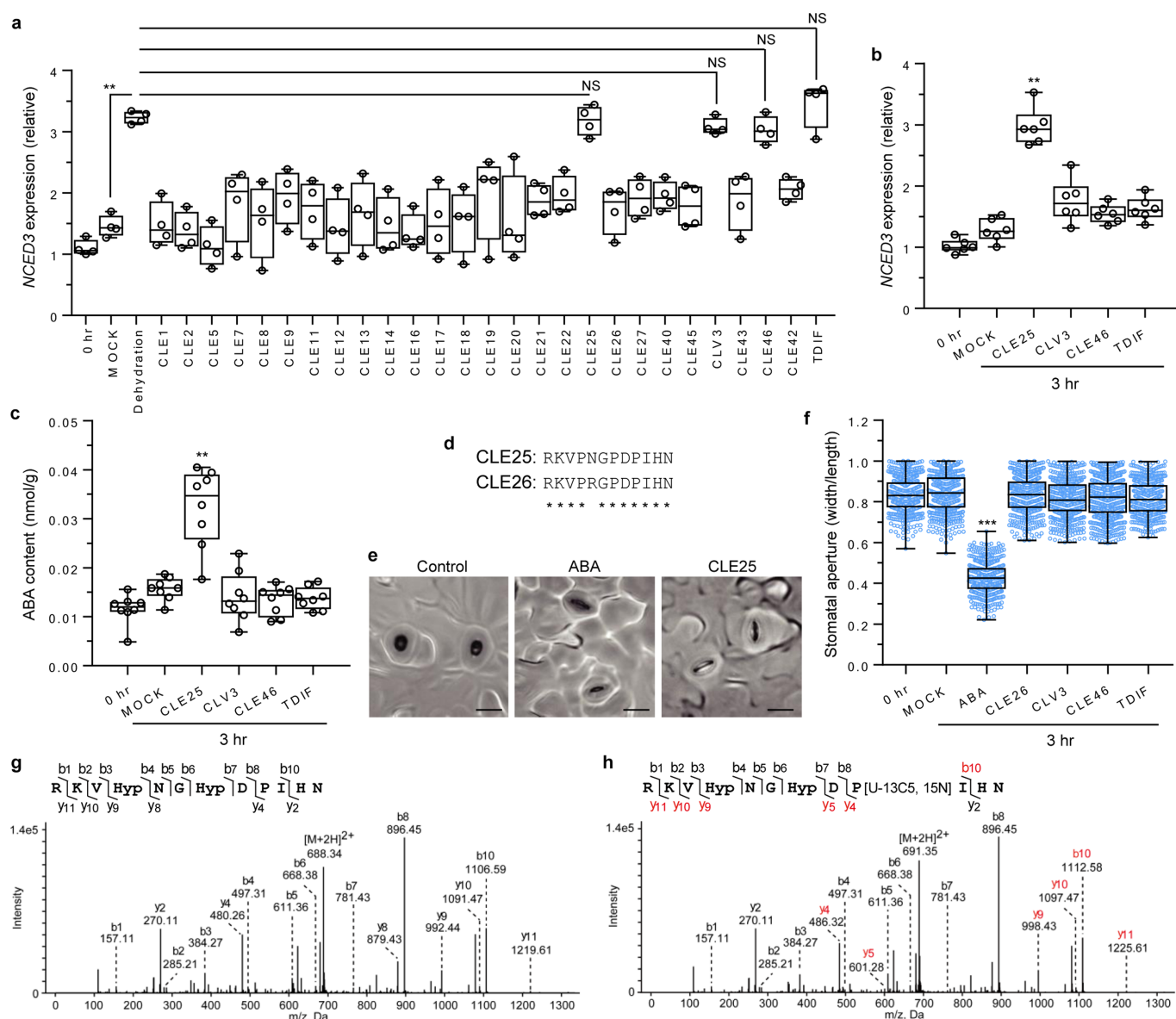
Extended Data Fig. 1 | Effects of synthetic CLE peptide application on NCED3 expression and stomatal closure in leaves, and movement of CLE25 peptide from roots to leaves. a, *NCED3* expression after application of 5 μM of each synthetic CLE peptide to roots for 3 h, or in response to dehydration stress for 3 h in leaves of wild-type plants (*n* = 4 biological replicates). *****P* < 0.01, no significant difference (NS) among treatment conditions as analysed by two-tailed Student's *t*-test (see Methods for exact *P* values). **b**, *NCED3* expression in leaves of wild-type plants after application of 1 μM peptide to roots for 3 h (*n* = 6 biological replicates). **c**, ABA content in leaves after application of 1 μM peptide to roots for 3 h (*n* = 8 biological replicates). **d**, Comparison of peptide sequences of CLE25 and CLE26. **e**, Typical images of wild-

type (*n* = 6 biological replicates), 1 μM-ABA- or 1 μM-CLE25-induced stomatal closure (*n* = 4 and 6 biological replicates, respectively). Scale bars, 20 μm. **f**, Roots of whole plants were incubated without (0 h on *x* axis, *n* = 547) or with 0.01% acetonitrile (mock, *n* = 519), ABA (*n* = 562) or each CLE peptide (*n* = 546, CLE26; *n* = 578, CLV3; *n* = 762, CLE46; *n* = 561, TDIF) for 3 h. Data are from three experiments. *****P* < 0.01, ******P* < 0.001 as analysed by one-way ANOVA followed by a Tukey's (**b**, **c**) or a Tukey–Kramer (**f**) post hoc test. **g**, **h**, Detection of non-labelled (**g**) and isotope-labelled (**h**) CLE25 peptide by nLC–MS/MS. These experiments were repeated two times independently with similar results. Hyp, hydroxyproline.
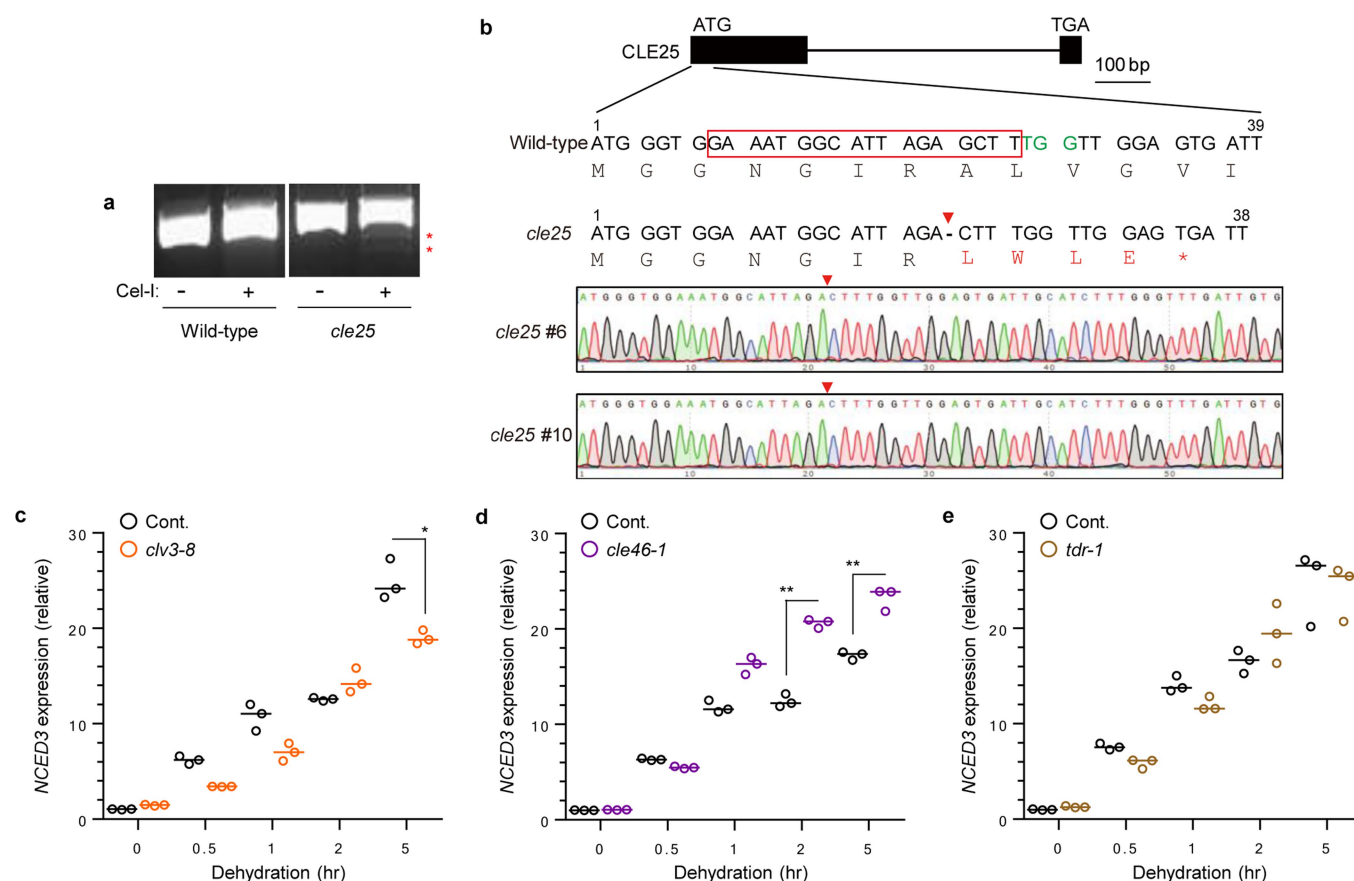
Extended Data Fig. 2 | The *cle25* mutants are generated using CRISPR–Cas9 method, and *clv3-8*, *cle46-1* and *tdr-1* mutants do not exhibit repression of *NCED3* expression in response to dehydration stress.
a, CEL-I analysis of T1 *cle25* mutants. The *CLE25* locus was amplified in wild-type and *cle25* mutants, then digested with CEL-I. The asterisk indicates mutated bands digested with CEL-I nuclease. These experiments were repeated four times independently with similar results. For gel source data, see Supplementary Fig. 1. **b**, CRISPR–Cas9-induced mutation detected by amplicon sequencing in T3 plants. The exons (boxes) and intron (line) indicate the schematic arrangement of the *CLE25* gene. Selected target sequences (18 base pairs) are shown in red boxes and protospacer adjacent motif sequences are shown as green characters.

A base deletion (guanosine at position 22) was detected in the genomic DNA of the mutants. The red triangle shows the position of the base deletion site. This mutation created a stop codon after the mutation site (asterisk in the amino acid sequence). (**c–e**, Dehydration-induced *NCED3* expression in *clv3-8* (**c**), *cle46-1* (**d**) and *tdr-1* (**e**) mutants was not repressed compared with that in wild-type (Cont.) plants in response to dehydration stress ($n = 3$ pooled biological replicates). CLV3, CLE46 and TDIF peptides do not have a primary function in the dehydration stress response that mediates ABA signalling. *$P < 0.05$, **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test ((**c–e**). The *clv3-8* mutant was a point mutant of CLV3. The *cle46-1* and *tdr-1* mutants were transfer DNA mutants of *CLE46* and *TDIF RECEPTOR*, respectively.
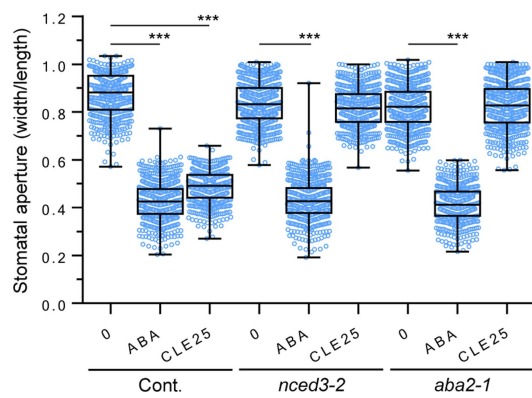
Extended Data Fig. 3 | *nced3-2* and *aba2-1* mutants treated with CLE25 peptide do not exhibit stomatal closure. Detached rosette leaves were incubated without (labelled '0' on *x* axis: $n = 505$, wild type (Cont.); $n = 647$, *nced3-2*; $n = 564$, *aba2-1*) or with ABA ($n = 617$, wild type; $n = 591$, *nced3-2*; $n = 467$, *aba2-1*) or the CLE25 peptide ($n = 505$, wild type; $n = 517$, *nced3-2*; $n = 570$, *aba2-1*) for 3 h. Data are from three experiments. ***$P < 0.001$ as analysed by one-way ANOVA followed by a Tukey–Kramer post hoc test.
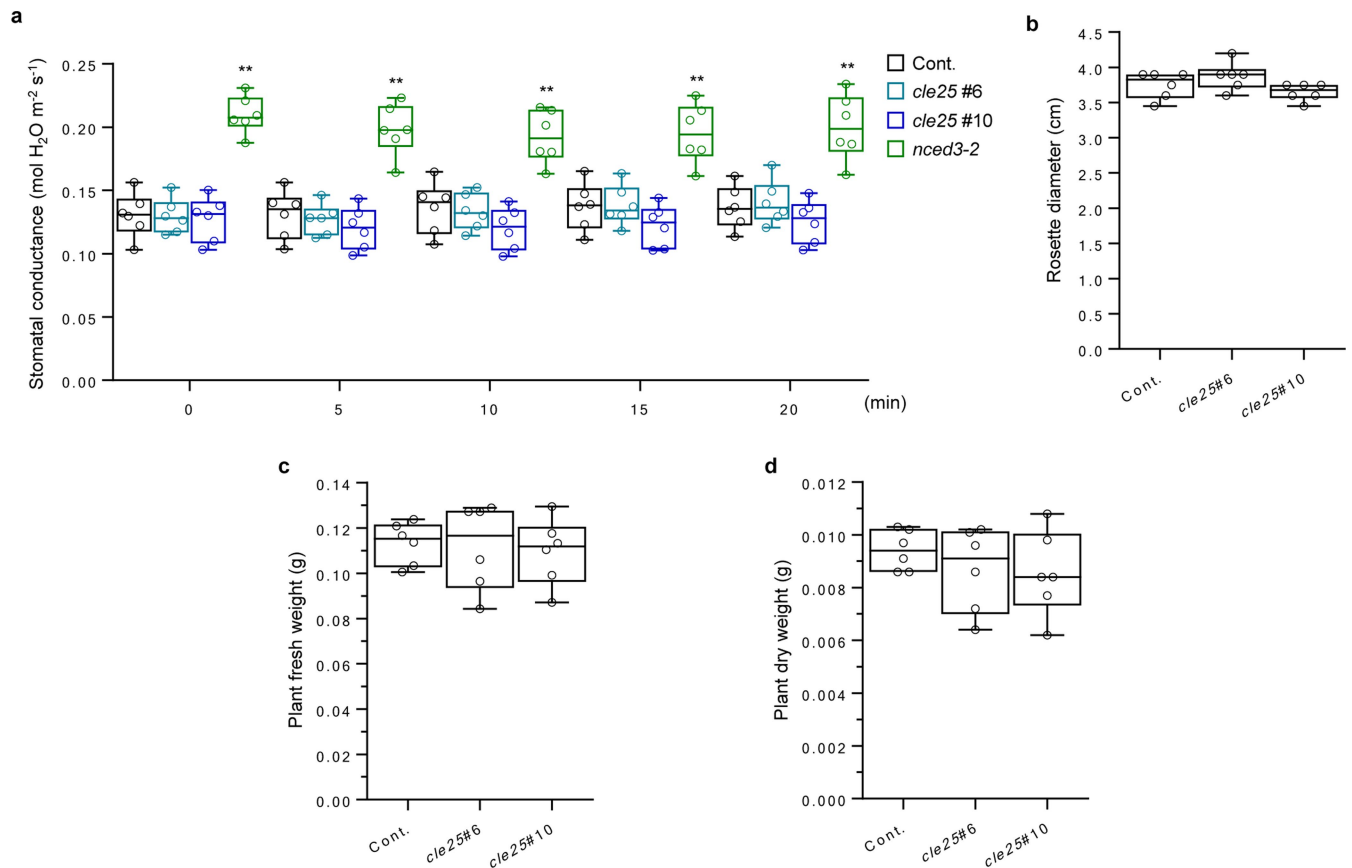
Extended Data Fig. 4 | Stomatal conductance, rosette diameter, fresh weight and dry weight of wild-type and *cle25* mutants grown on soil under control conditions. a, The stomatal conductance of wild-type (Cont.) plants, and *cle25* and *nced3-2* mutants ($n = 6$ biological replicate) was measured. Data are plotted at 5-min intervals for 20 min under control conditions. **b**, Rosette size of wild-type plants and *cle25* mutants ($n = 6$ biological replicates) grown on soil was scored. **c**, Fresh weight of wild-type plants and *cle25* mutants ($n = 6$ biological replicates) grown on soil was measured. **d**, Dry weight of wild-type plants and *cle25* mutants ($n = 6$ biological replicates) grown on soil was measured. **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test (**a–d**).
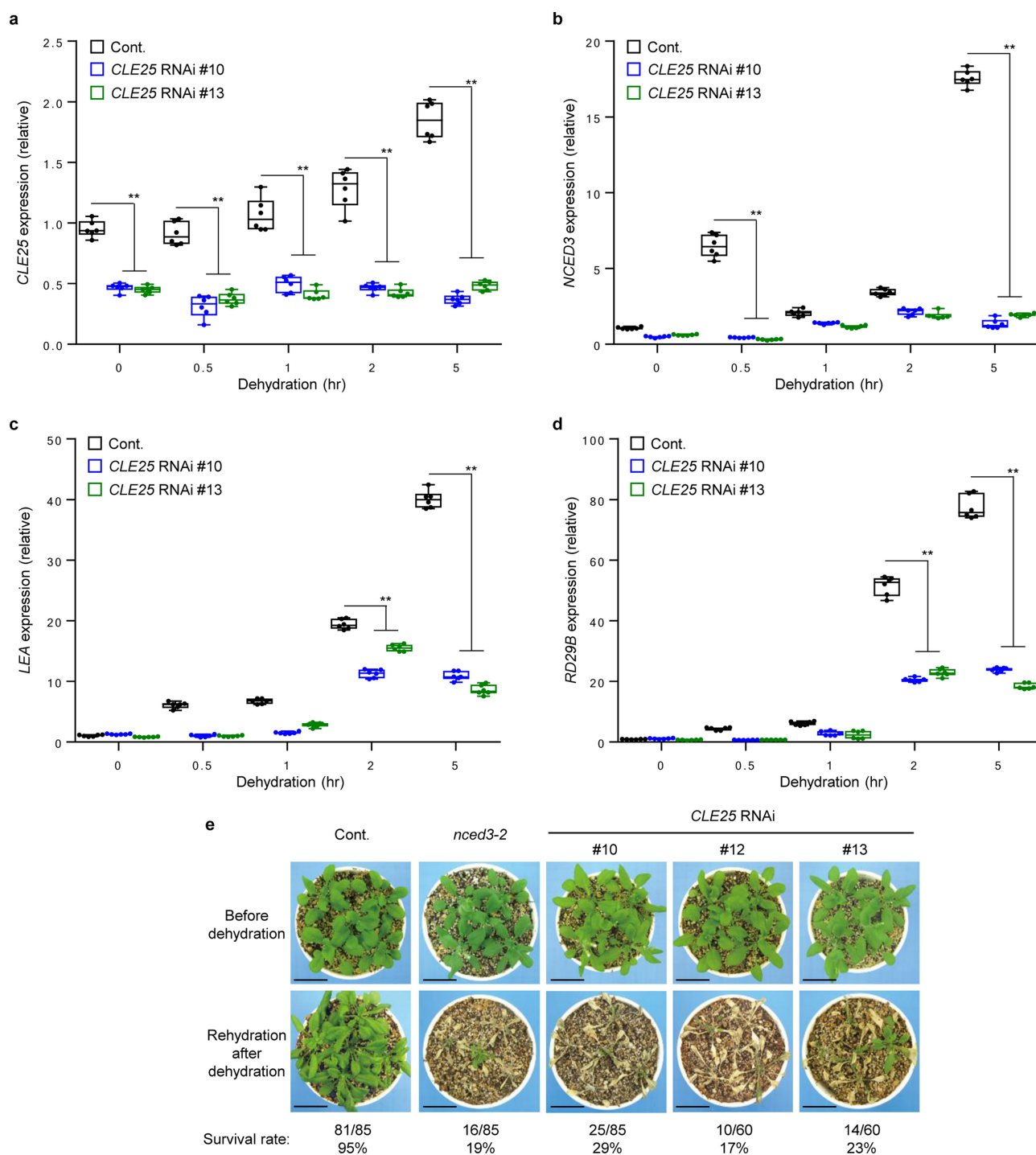
Extended Data Fig. 5 | Repression of *CLE25* in transgenic plants affects expression of dehydration-induced genes and hypersensitivity to dehydration stress. a–d, Dehydration-induced *CLE25* (**a**), *NCED3* (**b**), *LEA* (**c**) and *RD29B* (**d**) expression in wild-type (Cont.) and *CLE25* RNAi plants in response to dehydration stress (*n* = 6 biological replicates). *******P* < 0.01 as analysed by one-way ANOVA followed by a Tukey's post hoc test (**a–d**). **e,** *CLE25* RNAi plants and the *nced3-2* mutant had a dehydration stress-sensitive phenotype (plants per group; *n* = 85, wild type (Cont.); *n* = 85, *nced3-2*; *n* = 85, *CLE25* RNAi #10; *n* = 60, *CLE25* RNAi #12; *n* = 60, *CLE25* RNAi #13). Data are from three experiments. Scale bars, 2 cm.

a

b

R K V Hyp N G Hyp D P I H N

endogenous CLE25
MH^{3+}=459.24

synthetic CLE25
MH^{3+}=459.24

c

Accession	Description	Accumulation in T87 cells		Accumulation in liquid culture medium	
		Control (x 10^6)	Mannitol (x 10^6)	Control (x 10^6)	Mannitol (x 10^6)
ATCG00490	ribulose-bisphosphate carboxylases	4500	5300	125	88
AT3G55440	triosephosphate isomerase	3500	2800	N.D.	N.D.
AT2G36530	enolase	2800	2450	12.5	9.3
AT1G35720	annexin 1	2250	2150	N.D.	N.D.
AT3G44310	nitrilase 1	1950	2050	N.D.	N.D.
AT3G49120	peroxidase CB	1020	1550	6.2	13.5
AT1G67090	ribulose bisphosphate carboxylase small chain 1A	1850	1550	27.5	3.4
AT4G25100	Fe superoxide dismutase 1	1650	1450	58.5	66.5
AT4G13940	S-adenosyl-L-homocysteine hydrolase	1650	1400	N.D.	N.D.
AT5G22650	histone deacetylase 2B	1450	1350	N.D.	N.D.

Extended Data Fig. 6 | Endogenous CLE25 peptide is secreted extracellularly. *Arabidopsis* T87 cells were cultured with or without 0.4 M mannitol for 4 h. Then, peptides in the liquid culture medium were purified and analysed by nLC–MS/MS. **a**, Selected MS/MS ion chromatograms of the y4-ion from triply charged CLE25 peptides treated with 0.4 M mannitol. **b**, MS/MS spectra of endogenous (upper, with 0.4 M mannitol treatment) and synthetic (lower) CLE25 peptide obtained by nLC–MS/MS. Endogenous CLE peptide with 0.4 M mannitol treatment was detected only in di-hydroxy form. These experiments were repeated two times independently with similar results (**a**, **b**). Hyp, hydroxyproline. **c**, List indicates top 10 proteins in T87 cells in response to mannitol treatment. Amounts of these top 10 proteins accumulated in the liquid culture medium were the same under control conditions and in response to mannitol treatment. Cell lysis did not occur in response to mannitol treatment.

Extended Data Fig. 7 | CLE25 peptide moves from roots to leaves and modulates *NCED3* expression in leaves according to grafting experiments. a, **b**, *CLE25* expression (**a**) and *NCED3* expression (**b**) after application of CLE25 peptide to roots in leaves of wild-type (Cont.) and *CLE25* RNAi plants ($n = 6$ biological replicates). **c**, Dehydration-induced *CLE25* expression in grafted plants in which shoots and roots were grafted between wild-type and *CLE25* RNAi plants ($n = 6$ biological replicates).

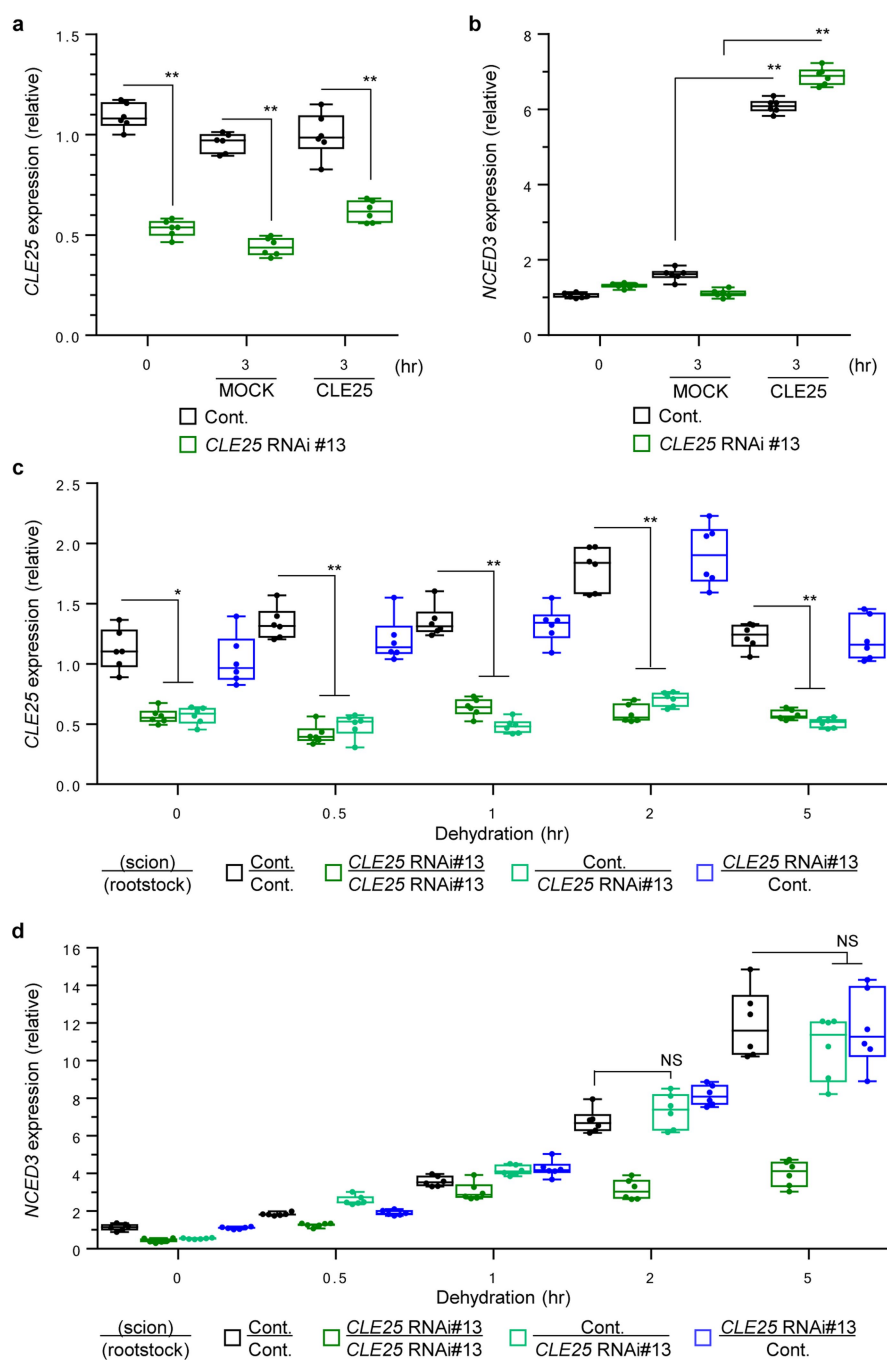*$P < 0.05$, **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test (**a**–**c**). **d**, Dehydration-induced *NCED3* expression in grafted leaves in which shoots and roots were grafted between wild-type and *CLE25* RNAi plants ($n = 6$ biological replicates). No significant difference (NS) among the three genotypes as analysed by two-tailed Student's *t*-test (see Methods for exact *P* values).
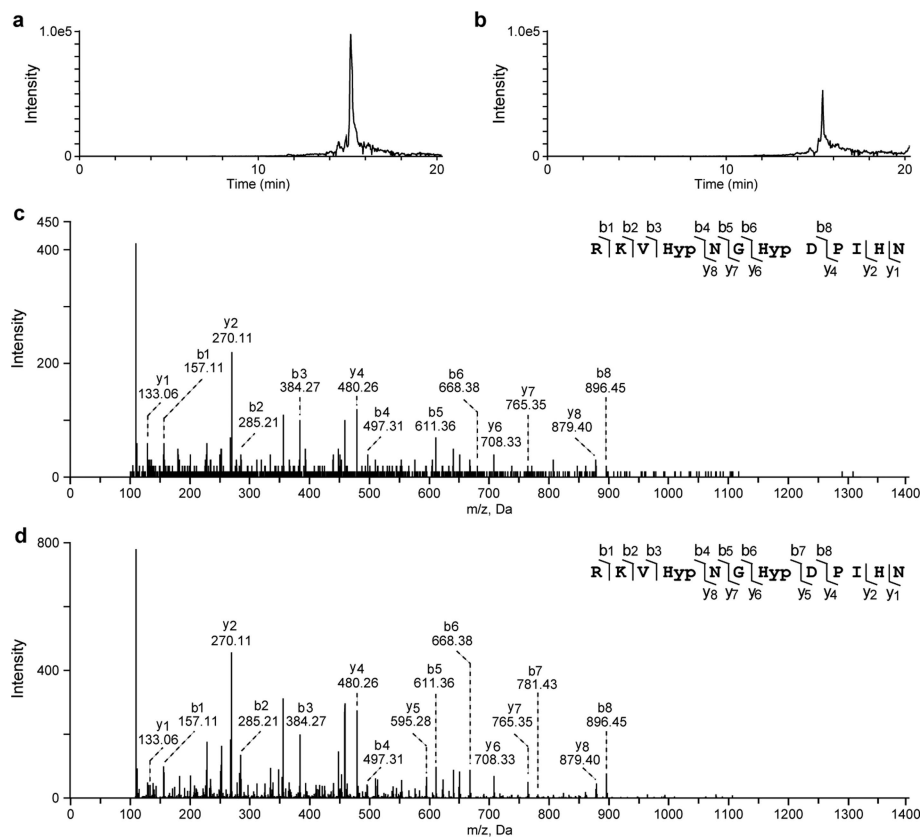
Extended Data Fig. 8 | Root-derived endogenous CLE25 peptide accumulates in dehydrated leaves. Accumulation of CLE25 peptide in leaves of wild-type and *cle25*-mutant shoot scions grafted onto wild-type rootstocks was analysed by nLC–MS/MS. **a**, **b**, MS/MS ion chromatograms of triple-charged CLE25 peptides in dehydrated leaves of grafted wild-type/wild-type (**a**) or *cle25* #10/wild-type (**b**) plants under 3-h dehydration-stress conditions. **c**, MS/MS spectra of endogenous CLE25 peptide in leaves of grafted *cle25* #10/wild-type plants under 3-h dehydration-stress conditions. **d**, MS/MS spectra of synthetic CLE25 peptide. These experiments were repeated two times independently with similar results (**a**–**d**). Hyp, hydroxyproline.
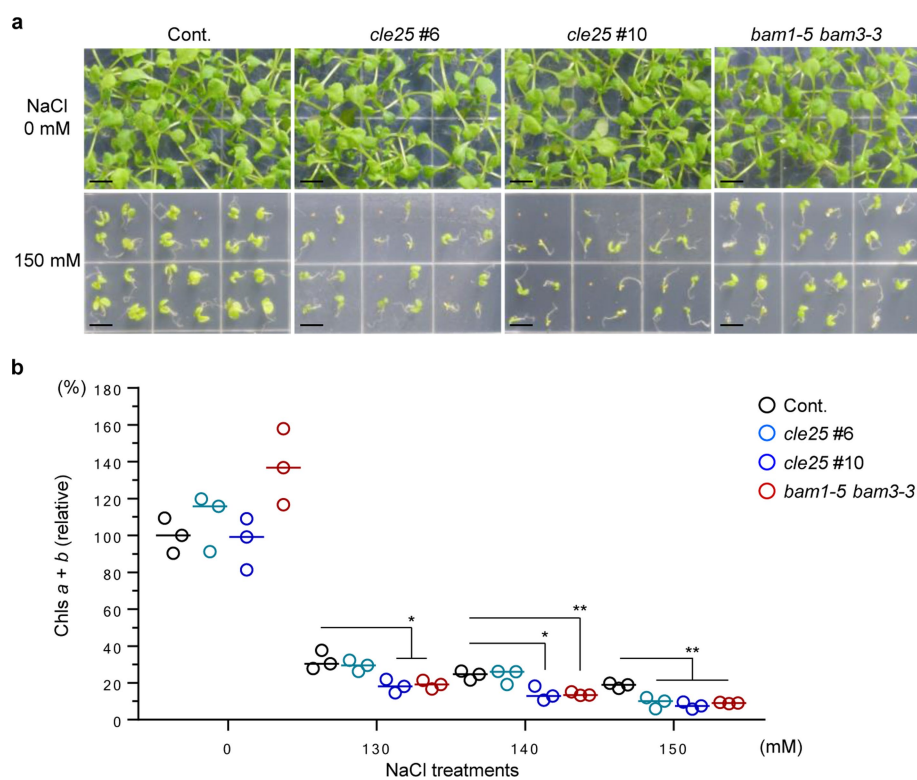
Extended Data Fig. 9 | cle25 and bam1-5 bam3-3 mutants show salinity stress-sensitive phenotype. a, Images represent 16-day-old seedlings for each genotype grown on germination medium agar plates containing 0 mM or 150 mM NaCl ($n = 3$ biological replicates). Scale bars, 0.5 cm. **b**, Measurements of relative chlorophyll contents in response to treatment with different NaCl concentrations were shown for 16-day-old seedlings after germination ($n = 3$ pooled biological replicates). *$P < 0.05$, **$P < 0.01$ as analysed by two-tailed Student's t-test (see Methods for exact P values).
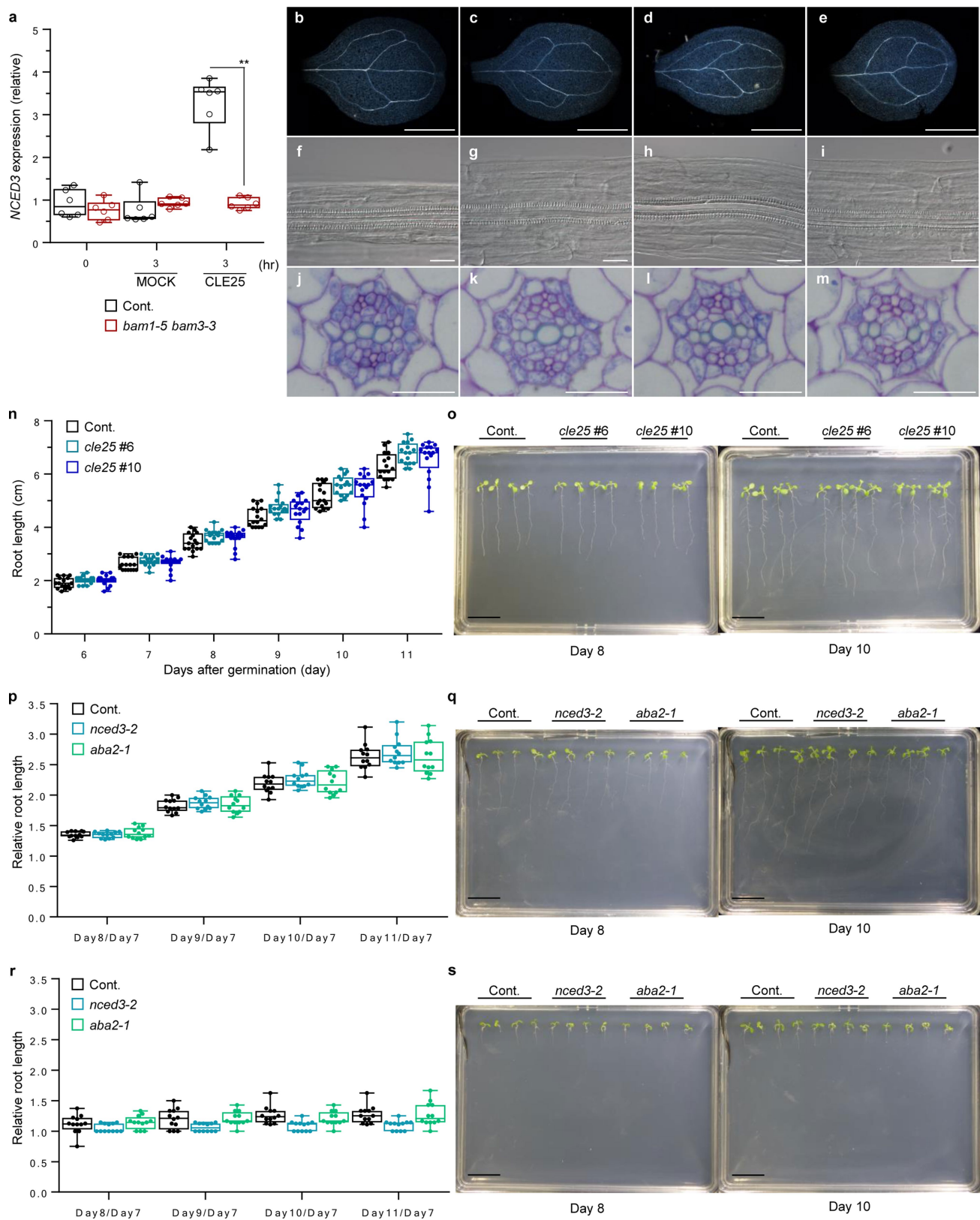
Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | *NCED3* expression in CLE25-treated leaves of *bam1-5 bam3-3* mutants, vascular development of wild-type, *cle25* and *bam1-5 bam3-3* mutants under control conditions and root growth phenotypes of *cle25*, *nced3-2* and *aba2-1* mutants under control conditions or long-term application of CLE25 peptide. a, *NCED3* expression in the leaves of wild-type (Cont.) and *bam1-5 bam3-3* mutants ($n = 6$ biological replicates) after application of CLE25 peptide to leaves. **$P < 0.01$ as analysed by one-way ANOVA followed by a Tukey's post hoc test. **b–e,** Microscopy images of the leaf vasculature of wild-type (**b**; $n = 6$ biological replicates), *cle25* #6 (**c**; $n = 12$ biological replicates), *cle25* #10 (**d**; $n = 12$ biological replicates) and *bam1-5 bam3-3* mutants (**e**; $n = 12$ biological replicates). Scale bars, 1 mm. **f–i,** Microscopy images of the protoxylem and metaxylem vessel formation of wild-type (**f**; $n = 4$ biological replicates), *cle25* #6 (**g**; $n = 3$ biological replicates), *cle25* #10 (**h**; $n = 4$ biological replicates) and *bam1-5 bam3-3* mutants (**i**; $n = 4$ biological replicates). Scale bars, 50 μm. **j–m,** Cross section of primary root of wild-type (**j**; $n = 9$ biological replicates), *cle25* #6 (**k**; $n = 10$ biological replicates), *cle25* #10 (**l**; $n = 10$ biological replicates) and *bam1-5 bam3-3* mutants (**m**; $n = 8$ biological replicates). Scale bars, 20 μm.

n, Root length of wild-type and *cle25* mutants from six to eleven days after germination, on germination medium agar plates ($n = 16$ biological replicates). **o,** Images represent 8-day-old or 10-day-old seedlings after germination of each genotype on germination medium agar plates ($n = 4$ biological replicates). Scale bars, 2 cm. **p,** Relative root length of wild-type, and *nced3-2* and *aba2-1* mutants from seven to eleven days after germination, on germination medium agar plate ($n = 12$ biological replicates). **q,** Images represent 8-day-old or 10-day-old seedlings after germination of each genotype on germination medium agar plates ($n = 3$ biological replicates). Scale bars, 2 cm. **r,** Relative root length of wild-type, and *nced3-2* and *aba2-1* mutants from seven to eleven days after germination, on germination medium agar plates containing 1 μM CLE25 peptide ($n = 12$ biological replicates). **s,** Images represent 8-day-old or 10-day-old seedlings after germination of each genotype on germination medium agar plates containing 1 μM CLE25 peptide ($n = 3$ biological replicates). Scale bars, 2 cm. Two-way ANOVA followed by a Tukey's post hoc test indicated that there were no differences among each genotype (**n, p, r**).

LETTER

Advanced maturation of human cardiac tissue grown from pluripotent stem cells

Kacey Ronaldson–Bouchard[1], Stephen P. Ma[1], Keith Yeager[1], Timothy Chen[1], LouJin Song[2], Dario Sirabella[1], Kumi Morikawa[2], Diogo Teles[1,3,4], Masayuki Yazawa[2] & Gordana Vunjak–Novakovic[1,5]*

Cardiac tissues generated from human induced pluripotent stem cells (iPSCs) can serve as platforms for patient-specific studies of physiology and disease[1–6]. However, the predictive power of these models is presently limited by the immature state of the cells[1,2,5,6]. Here we show that this fundamental limitation can be overcome if cardiac tissues are formed from early-stage iPSC-derived cardiomyocytes soon after the initiation of spontaneous contractions and are subjected to physical conditioning with increasing intensity over time. After only four weeks of culture, for all iPSC lines studied, such tissues displayed adult-like gene expression profiles, remarkably organized ultrastructure, physiological sarcomere length (2.2 μm) and density of mitochondria (30%), the presence of transverse tubules, oxidative metabolism, a positive force–frequency relationship and functional calcium handling. Electromechanical properties developed more slowly and did not achieve the stage of maturity seen in adult human myocardium. Tissue maturity was necessary for achieving physiological responses to isoproterenol and recapitulating pathological hypertrophy, supporting the utility of this tissue model for studies of cardiac development and disease.

Even the best available methods have limited ability to emulate the physiology of adult myocardium[1–12]; excitation–contraction coupling (requiring transverse tubules (T-tubules)), positive force–frequency relationship (requiring mature calcium handling) and efficient energy conversion (requiring oxidative metabolism) are notably absent[2,3,5,6,8–10]. Adult ventricular myocytes are uniquely organized for beating function, having densely packed sarcomeres, mitochondria, transverse tubules and sarcoplasmic or endoplasmic reticulum (SR/ER). Their mitochondria are positioned adjacent to sarcomeres and calcium pumps to enhance ATP diffusion; the sarcoplasmic reticulum provides fast delivery of stored calcium ions to contractile proteins; and the T-tubules synchronize heartbeats by concentrating L-type calcium channels, which are positioned close to the ryanodine receptors that release calcium ions from the SR/ER[13]. This highly specialized machinery for excitation–contraction coupling is not present in the fetal heart, but emerges after birth[14] with the switch from glycolytic to oxidative metabolism that supports the energy demands of the postnatal heart[15].

Human iPSC-derived cardiomyocytes (hiPS-CMs) can be matured by long-term culture and electrical, hydrodynamic and mechanical stimulation[8,9,12,16–18]. Recent studies have indicated that this in vitro maturation may not follow the in vivo developmental paradigm; high stimulation frequencies benefit maturation in vitro[9], whereas the native heart beats more slowly following birth[5,14]. We investigated the reasons why current strategies fail to develop the characteristics of adult myocardium. Because the responsiveness of hiPS-CMs to physical stimuli declines as differentiation progresses, we suggested that electromechanical conditioning should be initiated early, during the period of high cell plasticity. As the heart matures in response to energy demands, we further hypothesized that increasing the intensity of induced contractions would enhance the development of mature ultrastructure and function.

To test these hypotheses, we studied the maturation of human cardiac tissues grown from early-stage hiPS-CMs (day 12, immediately following the first spontaneous contractions) or late-stage hiPS-CMs (day 28, matured in culture). Cardiac tissues were assembled in a modular tissue platform that enabled individual control of the culture environment and physical signalling. hiPS-CMs (derived from three donors) and supporting fibroblasts were incorporated into fibrin hydrogel stretched between two flexible pillars (designed to provide mechanical forces similar to those in native myocardium) and subjected to electrical stimulation to induce auxotonic contractions. Three conditioning regimes were applied: (i) control (no stimulation); (ii) constant (three weeks at 2 Hz); and (iii) intensity training (two weeks at a frequency increasing from 2 Hz to 6 Hz by 0.33 Hz per day, followed by one week at 2 Hz). The resulting tissues were 6 mm long and 1.8 mm in diameter, and were evaluated in real time (for contractile and conductive behaviour and calcium handling) and by end-point assays (for gene expression, proteins and ultrastructure), using human fetal cardiac tissues (FCTs) and adult human heart ventricles as benchmarks (Fig. 1a, Extended Data Fig. 1a–e). Intensity-trained tissues grown from early-stage hiPS-CMs (hereafter early-stage intensity-trained) exhibited compact and well-differentiated cardiac muscle (Extended Data Fig. 1f–p) and marked changes in the expression of genes associated with adult-like conduction (increased *ITPR3*, *KCNH2*, decreased *HCN4*), maturation (increased *NPPB*, *MAPK1*, *PRKACA*), ultrastructure (increased *MYH7*, *GJA1*, *TNNI3*, *AKAP6*, *GJA5*, *JPH2*), energetics (increased *AKAP1*, *TFAM*, *PPARGC1A*) and calcium handling (increased *CAV3*, *BIN1*, *ATP2A2*, *RYR2*, *ITPR3*). The other early-stage-derived tissues, all late-stage-derived tissues and FCTs displayed immature cardiac phenotypes (Fig. 1b, Extended Data Fig. 2a, b).

Seeding with early-stage hiPS-CMs was critical for the response of the mature tissues to physical signals. Only the early-stage intensity-trained tissues displayed orderly signal propagation and anisotropic gap junctions. Among all tested groups, early-stage intensity-trained tissues had electrophysiological properties that were comparable to Biowires[9], including the shape of the action potential with its characteristic notch, the resting membrane potential of -70.0 ± 2.7 mV, the I_{K1} current (peak inward density of -9.9 ± 3.8 pA pF^{-1} and peak outward density of 0.30 ± 0.12 pA pF^{-1}) and the conduction velocity (25.0 ± 0.9 cm s^{-1}) (Fig. 1c, d, Extended Data Figs. 2, 3a–f, Supplementary Videos 1, 2).

Early-stage intensity-trained tissues also exhibited a positive force–frequency relationship (FFR), a hallmark of maturation not seen in other in vitro myocardial tissue models[5,6]. The generated forces markedly exceeded those in all other tested groups and FCT (Fig. 1f), but remained below those in adult myocardium[19] (44 mN mm^{-2}). Directly measured forces and contraction amplitudes increased approximately twofold over the range of stimulation frequencies (1–6 Hz) during the maturation of early-stage intensity-trained tissues, indicating the maturation of contractile behaviour. These tissues acquired regular contraction profiles, in contrast to late-stage intensity-trained tissues

[1]Laboratory for Stem Cells and Tissue Engineering, Department of Biomedical Engineering, Columbia University, New York, NY, USA. [2]Department of Rehabilitation and Regenerative Medicine, Department of Pharmacology, College of Physicians and Surgeons, Columbia University, New York, NY, USA. [3]Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal. [4]ICVS/3B's, PT Government Associate Laboratory, Braga–Guimarães, Portugal. [5]Department of Medicine, Columbia University, New York, NY, USA. *e-mail: gv2131@columbia.edu
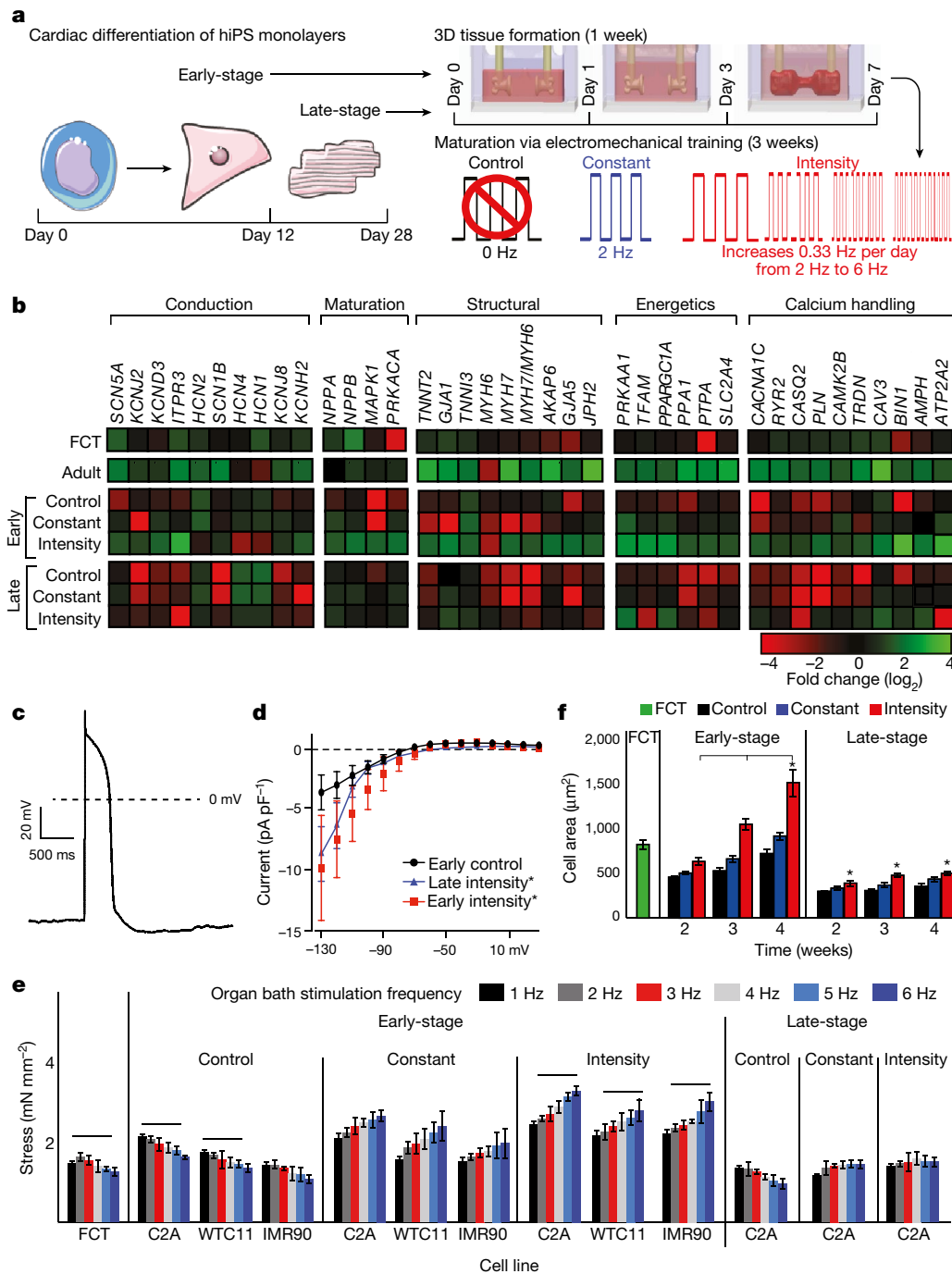
Fig. 1 | Intensity training of cardiac tissues derived from early-stage hiPS-CMs enhances maturation. a, Experimental design: early-stage or late-stage hiPS-CMs and supporting fibroblasts were encapsulated in fibrin hydrogel to form tissues stretched between two elastic pillars and made to contract by electrical stimulation. Gradual increase in frequency of stimulation to supra-physiological levels (intensity regime) was compared to stimulation at constant frequency (constant regime), unstimulated controls and human adult and fetal heart ventricles. **b**, Gene expression data for six groups of cardiac tissues, and adult and fetal heart ventricles. **c**, Action potential for the early-stage intensity-trained group. **d**, I_{K1} current–voltage ($I–V$) curves (mean ± s.d.). **e**, Early-stage

intensity-trained tissues from all three iPSC lines (C2A, WTC11, IMR90), but not the other groups, developed a positive force–frequency relationship after four weeks of culture. Line above graph indicates $P < 0.05$ for the 2–6 Hz group versus other training regimes using two-way ANOVA followed by Tukey's honest significant difference (HSD) test. **f**, Cell area over time. Line above graph indicates $P < 0.05$ versus other timepoints using two-way ANOVA followed by Tukey's HSD test; *$P < 0.05$ versus control group using one-way ANOVA followed by Tukey's HSD test. Data in **e** and **f** are mean ± 95% confidence interval (CI). Sample sizes are shown in Supplementary Information, 'Main figure data sample sizes'.

(Extended Data Figs. 3g–i, 4a–d, Supplementary Videos 1, 2). The surrogate measurements of force from calcium recordings in early-stage intensity-trained tissues (Extended Data Fig. 3j, k), were consistent with the direct force measurements.

Cell populations were dominated by cardiomyocytes, and the MLC2v$^+$:MLC2a$^+$ ratio, an indicator of cardiomyocyte maturity, depended on the stimulation regime and developmental stage of

hiPS-CMs. The increasing contractile demands induced the adult-like cardiac morphology that is necessary for high force generation in early-stage intensity-trained tissues. The cell size increased (an indicator of physiological hypertrophy[8]) and both cells and nuclei were elongated (an indicator of maturation[8]). The sarcomere length reached 2.2 μm, a similar value to that of adult human ventricular myocytes[8]. The contractile capacity, fraction of cells containing sarcomeres and
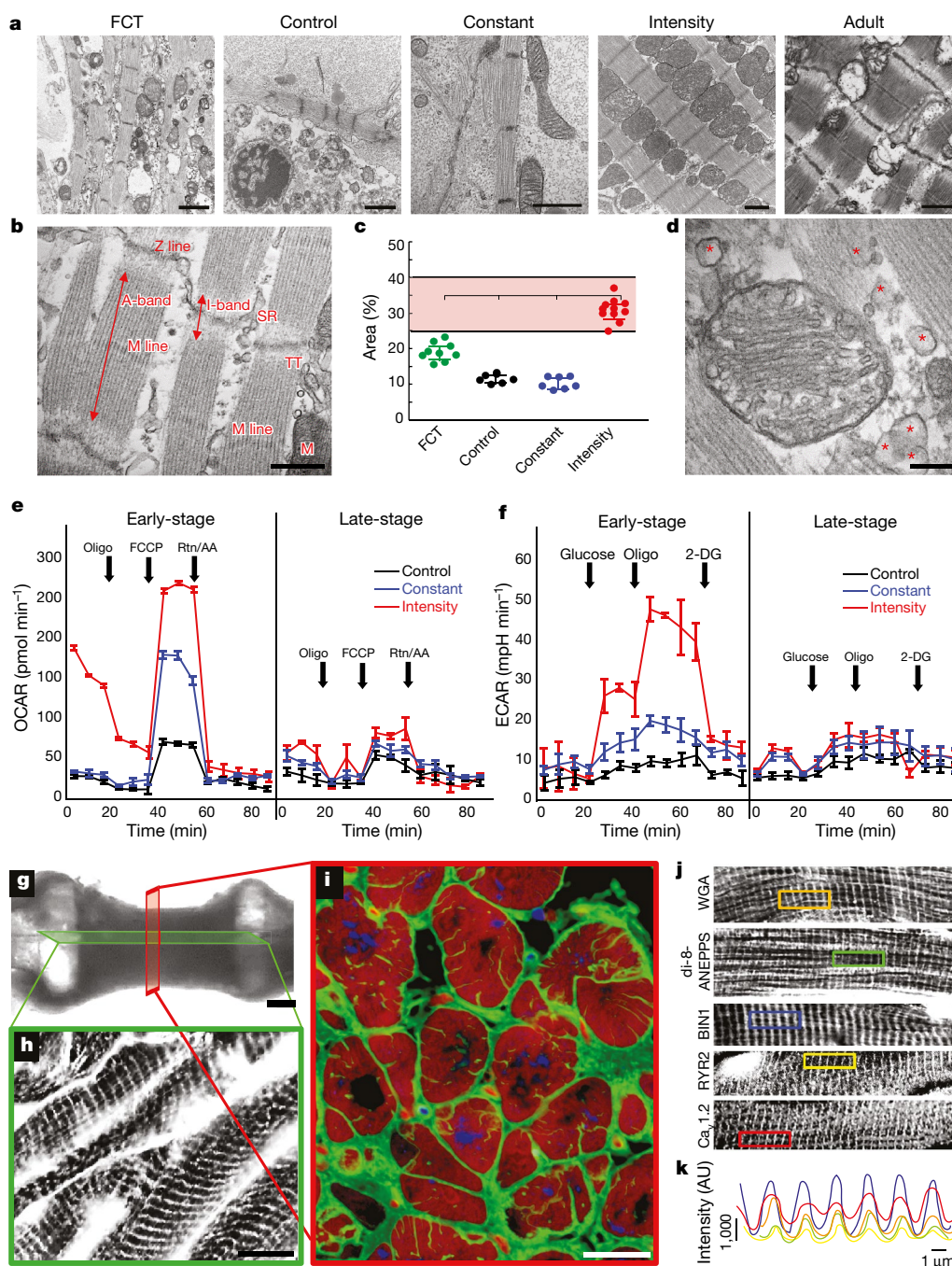
Fig. 2 | Enhanced cardiac ultrastructure, bioenergetics and T-tubule formation in early-stage intensity-trained tissues derived from C2A cells. a, Transmission electron microscopy (TEM) of tissues and cardiac tissue models. Scale bars, 1 μm. **b**, **d**, **g–k**, Early-stage intensity-trained tissues cultured for four weeks and derived from C2A cells. **b**, Registers of sarcomeres, showing A-bands, I-bands, M lines, Z lines, sarcoplasmic reticulum (SR) and T-tubules (TT). Scale bar, 1 μm. **c**, Density of mitochondria; shaded area represents range of values measured in adult human heart. Line above graph indicates $P < 0.05$ versus other training regimes using two-way ANOVA followed by Tukey's HSD test. **d**, Lipid droplets (red asterisk). Scale bar, 1 μm. **e**, Oxygen consumption rate (OCR). Oligo, oligomycin; FCCP, carbonyl cyanide-4-(trifluoromethoxy) phenylhydrazone; Rtn/AA, rotenone and antimycin A. **f**, Extracellular acidification rate. 2-DG, 2-deoxyglucose. **g–i**, Cross-sections taken to evaluate T-tubules: bright field view (**g**; scale bar, 500 μm); T-tubules (**h**, **i**; green, WGA; red, cardiac troponin T (cTnT); blue, nuclei; scale bar, 10 μm). **j**, Calcium handling ultrastructure. Scale bar, 15 μm. **k**, Regular spacing of calcium handling proteins shown in **j**. AU, arbitrary units. Data are mean ± 95% CI; sample sizes are shown in Supplementary

organization of sarcomeric α-actinin also resembled those of adult human myocardium (Fig. 1e, Extended Data Figs. 4e–k, 5).

Ultrastructural development was dependent on the stimulation regime and the developmental stage of hiPS-CMs from which the tissues were derived. Only early-stage intensity-trained tissues displayed orderly registers of sarcomeres with I-bands, A-bands, M lines, Z lines, desmosomes, intercalated discs, a high density of mitochondria

positioned adjacent to the contractile machinery and proteins organized for increased energetics (Fig. 2a, b, Extended Data Figs. 6, 7a, b).

Whereas the fetal heart favours glucose as the primary energy substrate[19], the increased workload in the postnatal heart results in mature mitochondria that are optimized for fatty acid oxidation[20,21]. The per cent area of mitochondria in early-stage intensity-trained tissues (30 ± 2.9%) was similar to those measured in adult human myocardium[22,23]. Active

biogenesis[24] was associated with the production of phospholipids near the sarcomeres, a switch to oxidative metabolism and the formation of T-tubules (Fig. 2c–f, Extended Data Fig. 7c, d).

Early-stage intensity-trained tissues contained robust T-tubules, both longitudinally and in cross-sections. T-tubules (measured using wheat germ agglutinin (WGA) and di-8-ANEPPS) were co-localized with the bridging integrator 1 (BIN1), ryanodine receptor 2 (RYR2), and L-type calcium channels (Ca$_V$1.2, encoded by *CACNA1C*) with spacing optimized for calcium handling (Fig. 2g–k, Extended Data Fig. 8), as in the adult heart[25]. These tissues displayed spatially uniform cell densities, presumably owing to the enhanced transport of nutrients and metabolites during tissue contractions, generated the highest force, and expressed the Ca^{2+}-induced Ca^{2+} release (CICR) modulators RYR2 (control of SR/ER calcium release) and BIN1 (control of ion flux along T-tubules[26]) (Fig. 3a, Extended Data Figs. 8a, f, g, 9a).

The frequency-dependent acceleration of relaxation (FDAR), an intrinsic property of adult myocardium that was observed for early-stage intensity-trained tissues showed that the tissues subjected to supra-threshold electrical pacing regimes developed mechanisms to respond to the increasing workload. The presence of ultrastructural machinery for contraction–relaxation was confirmed by the positioning of T-tubules in proximity to the cardiac calcium pump SERCA2A (encoded by *ATP2A2*) and the sodium–calcium exchanger NCX1 (encoded by *SLC8A1*). Consistently, transcription of the genes responsible for clearing cytosolic calcium (*ATP2A2* and *SLC8A1*[18]) increased over time, and the sequestration and extrusion of calcium became faster, enabling the hiPS-CMs to relax and respond to contractile triggers. Blocking Ca$_V$1.2 with nifedipine or verapamil gradually reduced calcium transients in a training-dependent manner, while the response to caffeine indicated that only the early-stage intensity-trained tissues had functional intracellular calcium stores. Blocking SERCA with thapsigargin treatment to prevent SR/ER calcium uptake halted calcium transients,

indicating that they are dependent on a functional sarcoplasmic reticulum. Subsequent addition of caffeine had no effect, consistent with calcium depletion of the SR/ER (Fig. 3b–d, Extended Data Fig. 9b–h).

Post-rest potentiation confirmed the functionality of SR/ER calcium stores in early-stage intensity-trained tissues. None of the other tested tissues responded to increased calcium levels or developed calcium alternans, owing to the lack of T-tubules and inefficient coupling between intracellular calcium entry and release. When CICR was blocked with ryanodine to test RYR2 function, only early-stage intensity-trained tissues showed a response, probably owing to the presence of T-tubules, which are necessary for CICR. Notably, the positive FFR was blunted by ryanodine treatment, and completely reversed when calcium sequestration by SERCA2a was blocked with thapsigargin, indicating the importance of both CICR and the reuptake of calcium into the sarcoplasmic reticulum (Fig. 3e, Extended Data Fig. 9i–k).

Because a functional β-adrenergic receptor system is dependent on both intracellular calcium reserves and the proximity of Ca$_V$1.2 channels and T-tubules[4,26,27], comprehensive responses to β-adrenergic agonists are an indicator of phenotypic maturation[28]. We investigated whether early-stage intensity-trained tissues had an ionotropic response to iso-proterenol, since this effect is not seen in current in vitro cardiac tissue models[2,10]. We detected positive chronotropic, ionotropic and lusitropic responses to isoproterenol in early-stage intensity-trained tissues, with EC$_{50}$ (half-maximum effective concentration) values corresponding to those observed in clinical studies[29] (Fig. 3f–h, Extended Data Fig. 10a, b).

Tissue maturity was necessary to recapitulate critical aspects of cardiac hypertrophy (HCM), a leading cause of sudden cardiac death in athletes[2]. As expected, hypertrophic tissues displayed decreased beating frequency and increased durations of intracellular calcium transients and decay times relative to healthy controls, and were not able to electromechanically capture when stimulated at high frequencies. The onset of HCM diminished the FDAR and resulted in a negative FFR, in
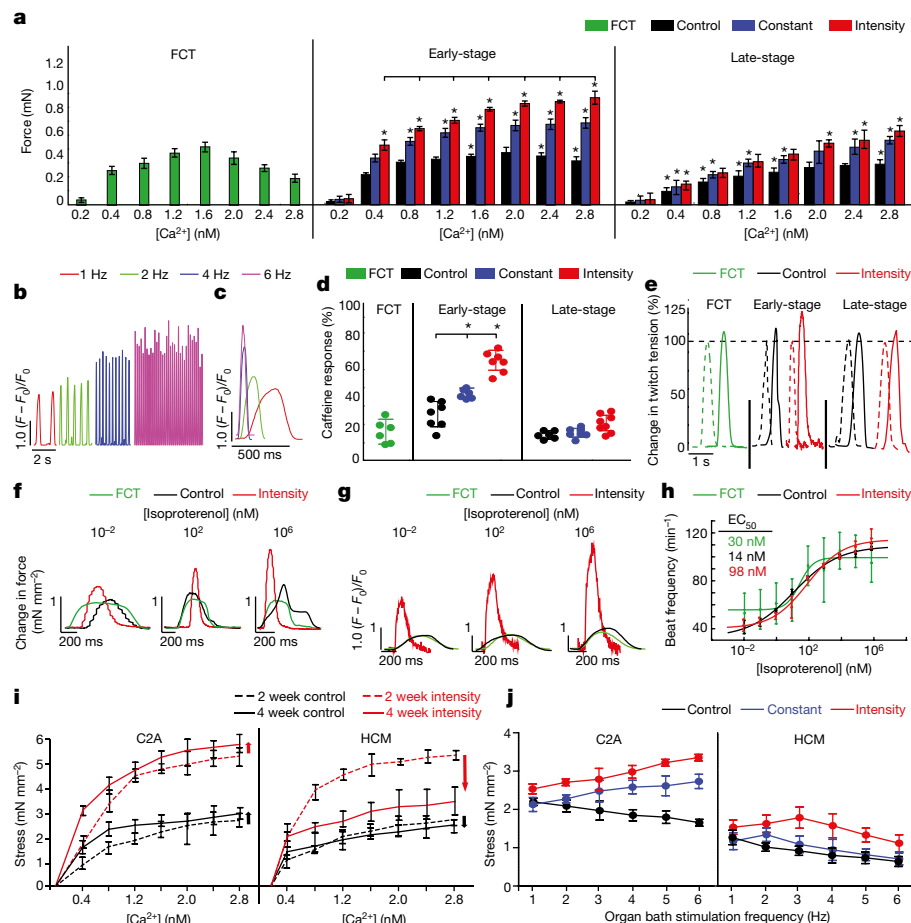
Fig. 3 | Mature calcium handling in early-stage intensity-trained tissues derived from C2A cells. a, Force of contraction during CICR. **b, c**, FDAR, shown by calcium traces. **d**, Calcium release after stimulation with 5 mM caffeine. **e**, Force traces during post-rest potentiation with 10 s of rest. **f–h**, Ionotropic (**f**), lusitropic (**g**) and chronotropic (**h**) dose-dependent responses. **i, j**, Cardiac tissue models of pathological hypertrophy (HCM) show decreased CICR over time (**i**) and reversal of positive FFR at higher pacing rates (**j**). Data are for C2A cell line after four weeks of culture; mean ± 95% CI. Sample sizes are shown in Supplementary Information, 'Main figure data sample sizes'. *$P < 0.05$ versus FCT using one-way ANOVA followed by Tukey's HSD test. In **a** and **d**, line above graph indicates $P < 0.05$ versus other training regimes using two-way ANOVA followed by Tukey's HSD test.

contrast to healthy tissues. Differences between healthy and diseased groups were most pronounced in the intensity-trained tissues (Fig. 3i, j, Extended Data Fig. 10c–f).

A recent study[12] reported the culture of large ($7 \times 7\,mm^2$ to $36 \times 36\,mm^2$) and thin ($50\,\mu m$) human heart tissues, grown without exogenous stimulation, that displayed less developed ultrastructure, no evidence of oxidative metabolism, slightly negative FFR, comparable APD and conduction velocity, and approximately fourfold higher generated force per unit cross-sectional tissue area when compared to the tissues cultured here. It would be instructive to explore how the different tissue geometries (very thin patches versus cylindrical muscle) and culture protocols (no external stimulation versus intensity training)[12] contributed to the measured differences in structural and functional tissue outcomes in comparison to the present study.

In summary, we have demonstrated that adult-like human cardiac tissue can be grown from hiPS-CMs in fibrin hydrogel subjected to stretch and auxotonic contractions in just four weeks of in vitro culture. Two methodological advances underlie the accelerated cardiac maturation: the formation of tissues from early-stage hiPS-CMs, which displayed marked plasticity immediately after the initiation of spontaneous contractions; and physical conditioning with increasing intensity (mimicking mechanical loading during the fetal–postnatal transition). Under these conditions, tissues developed adult-like gene expression and tissue ultrastructure throughout the tissue volume, oxidative metabolism, FDAR, positive FFR and physiological calcium handling.

A notable result of our study is that highly accelerated and extensive maturation of molecular, structural and metabolic features of cardiac tissue was associated with slower and less complete establishment of mature cardiac function. We have demonstrated that physiological cell density is not sufficient to achieve adult-like mechanical function; that FDAR and positive FFR can be established at subnormal levels of force generation; and that T-tubules and oxidative metabolism are required for physiological FFR and calcium handling. Our tissue model does not recapitulate the macroscopic structure of the myocardium, and the maturation period of four weeks may be too short to establish all the functional features of adult myocardium. These factors may contribute to the contrast between the impressive morphological maturation and the less complete functional maturation. It would therefore be instructive to use this human cardiac tissue model to study the progression of functional maturation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0016-3.

1. Bellin, M., Marchetto, M. C., Gage, F. H. & Mummery, C. L. Induced pluripotent stem cells: the new patient? *Nat. Rev. Mol. Cell Biol.* **13**, 713–726 (2012).
2. Matsa, E., Burridge, P. W. & Wu, J. C. Human stem cells for modeling heart disease and for drug discovery. *Sci. Transl. Med.* **6**, 239 (2014).
3. Wang, G. et al. Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat. Med.* **20**, 616–623 (2014).
4. Yazawa, M. et al. Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome. *Nature* **471**, 230–234 (2011).
5. Yang, X., Pabon, L. & Murry, C. E. Engineering adolescence: maturation of human pluripotent stem cell-derived cardiomyocytes. *Circ. Res.* **114**, 511–523 (2014).
6. Feric, N. T. & Radisic, M. Maturing human pluripotent stem cell-derived cardiomyocytes in human engineered cardiac tissues. *Adv. Drug Deliv. Rev.* **96**, 110–134 (2016).
7. Domian, I. J. et al. Generation of functional ventricular heart muscle from mouse ventricular progenitor cells. *Science* **326**, 426–429 (2009).
8. Lundy, S. D., Zhu, W. Z., Regnier, M. & Laflamme, M. A. Structural and functional maturation of cardiomyocytes derived from human pluripotent stem cells. *Stem Cells Dev.* **22**, 1991–2002 (2013).
9. Nunes, S. S. et al. Biowire: a platform for maturation of human pluripotent stem cell-derived cardiomyocytes. *Nat. Methods* **10**, 781–787 (2013).
10. Mannhardt, I. et al. Human engineered heart tissue: analysis of contractile force. *Stem Cell Reports* **7**, 29–42 (2016).
11. Ribeiro, M. C. et al. Functional maturation of human pluripotent stem cell derived cardiomyocytes in vitro—correlation between contraction force and electrophysiology. *Biomaterials* **51**, 138–150 (2015).
12. Shadrin, I. Y. et al. Cardiopatch platform enables maturation and scale-up of human pluripotent stem cell-derived engineered heart tissues. *Nat. Commun.* **8**, 1825 (2017).
13. Brette, F. & Orchard, C. T-tubule function in mammalian cardiac myocytes. *Circ. Res.* **92**, 1182–1192 (2003).
14. Wiegerinck, R. F. et al. Force frequency relationship of the human ventricle increases during early postnatal development. *Pediatr. Res.* **65**, 414–419 (2009).
15. Lopaschuk, G. D. & Jaswal, J. S. Energy metabolic phenotype of the cardiomyocyte during development, differentiation, and postnatal maturation. *J. Cardiovasc. Pharmacol.* **56**, 130–140 (2010).
16. Jackman, C. P., Carlson, A. L. & Bursac, N. Dynamic culture yields engineered myocardium with near-adult functional output. *Biomaterials* **111**, 66–79 (2016).
17. Radisic, M. et al. Functional assembly of engineered myocardium by electrical stimulation of cardiac myocytes cultured on scaffolds. *Proc. Natl Acad. Sci. USA* **101**, 18129–18134 (2004).
18. Eng, G. et al. Autonomous beating rate adaptation in human stem cell-derived cardiomyocytes. *Nat. Commun.* **7**, 10312 (2016).
19. Hasenfuss, G. et al. Energetics of isometric force development in control and volume-overload human myocardium. Comparison with animal species. *Circ. Res.* **68**, 836–846 (1991).
20. Chung, S. et al. Mitochondrial oxidative metabolism is required for the cardiac differentiation of stem cells. *Nat. Clin. Pract. Cardiovasc. Med.* **4**, S60–S67 (2007).
21. Gong, G. et al. Parkin-mediated mitophagy directs perinatal cardiac metabolic maturation in mice. *Science* **350**, aad2459 (2015).
22. Porter, G. A. Jr et al. Bioenergetics, mitochondria, and cardiac myocyte differentiation. *Prog. Pediatr. Cardiol.* **31**, 75–81 (2011).
23. Vega, R. B., Horton, J. L. & Kelly, D. P. Maintaining ancient organelles: mitochondrial biogenesis and maturation. *Circ. Res.* **116**, 1820–1834 (2015).
24. Gottlieb, R. A. & Bernstein, D. Metabolism. Mitochondria shape cardiac metabolism. *Science* **350**, 1162–1163 (2015).
25. Sun, R., Bouchard, M. B. & Hillman, E. M. C. SPLASSH: Open source software for camera-based high-speed, multispectral in-vivo optical image acquisition. *Biomed. Opt. Express* **1**, 385–397 (2010).
26. Hong, T. et al. Cardiac BIN1 folds T-tubule membrane, controlling ion flux and limiting arrhythmia. *Nat. Med.* **20**, 624–632 (2014).
27. Bers, D. M. Cardiac excitation–contraction coupling. *Nature* **415**, 198–205 (2002).
28. Huebsch, N. et al. Miniaturized iPS-cell-derived cardiac muscles for physiologically relevant drug response analyses. *Sci. Rep.* **6**, 24726 (2016).
29. Tulloch, N. L. et al. Growth of engineered human myocardium with mechanical loading and vascular coculture. *Circ. Res.* **109**, 47–59 (2011).

Author contributions K.R.-B. and G.V.-N. designed the study. K.R.-B., K.Y. and G.V.-N. designed the tissue culture platform. K.R.-B., S.P.M., T.C. and D.T. cultured all tissue types with different stimulation conditions, and performed real-time and end-point assessments of tissue properties (including gene expression, histomorphology, ultrastructure, distributions of cardiac proteins, contractile behaviour and calcium handling). K.R.-B., S.P.M., T.C. and D.T. independently replicated the entire process of tissue cultivation and assessment. D.S. and L.S. expanded iPSCs and derived cardiomyocytes. K.R.-B. performed immunostaining for the presence of T-tubules. M.Y., K.M. and L.S. conducted single-cell dissociation and electrophysiology experiments. K.R.-B., S.P.M., T.C., M.Y. and G.V.-N. interpreted data and wrote the manuscript.

METHODS

Cardiac differentiation of human iPSCs. Human induced pluripotent stem cells were obtained through material transfer agreements from S. Duncan, University of Wisconsin (C2A line), B. Conklin, Gladstone Institute (WT11 line) and M.Y., Columbia University (IMR90 line) and routinely checked for mycoplasma contamination. iPSCs were expanded on growth-factor-reduced Matrigel-coated plates (Corning) in mTeSR1 medium (Stemcell Technologies) that was changed on a daily basis, and passaged at 85–95% confluence in a 1:6 split using Accutase (Life Technologies). For the first 24 h after passaging, the culture medium was supplemented with 5 μM Y-27632 dihydrochloride (Tocris, 1254).

Cardiac differentiation of iPSCs was initiated in confluent monolayers by replacing the mTeSR1 medium with RPMI + B27−insulin medium, consisting of RPMI-1640 (Life Technologies), 1 × B27 supplement without insulin (a source of omega-3 fatty acids and the thyroid hormone that promotes cardiac maturation; Life Technologies), 100 U penicillin (Life Technologies), 0.1 mg/ml streptomycin (Life Technologies) and 50 μg/ml ascorbic acid (Sigma, A4544). During the first 24 h, the medium was further supplemented with activin A (50 ng/ml, R&D Systems) and bone morphogenetic protein 4 (BMP4, 25 ng/ml, R&D systems). From 24–72 h, the RPMI + B27−insulin medium was supplemented with vascular endothelial growth factor (VEGF, 10 ng/ml, R&D systems). Beyond 72 h through to the end of the differentiation process (up to 12 days), RPMI + B27 medium, consisting of RPMI-1640, 1 × B27 supplemented with insulin (Life Technologies), 100 U penicillin, 0.1 mg/ml streptomycin and 50 μg/ml ascorbic acid, was used and refreshed every two days. At day 12, the cells were characterized by flow cytometry using the cardiomyocyte-specific marker cTnT (clone 13-11, NeoMarkers). Differentiation typically resulted in cell populations containing 80–90% cTnT$^+$ cells at day 12, which were subsequently used in experiments without selection for cardiomyocytes.

Human fetal cardiac tissues. Fetal hearts were purchased as surgical waste from Advanced Bioscience Resources (Alameda, CA), and delivered on ice within 2.5 h of surgery. Left ventricles were sectioned from the apex towards the atria into 7 mm long × 2 mm wide strips, washed three times in Hank's Balanced Salt Solution (Gibco), transferred to low attachment six-well plates (Nunc) containing RPMI + B27 medium, and placed into the incubator for 1 h before taking measurements. FCT strips were analysed in a similar manner to the cardiac tissues for contractile behaviour, force generation, gene expression, cardiac proteins, ultrastructure and histomorphology, as detailed below. In addition, RNA isolated from 32 pooled fetal hearts, from gestational weeks (GW)21–37, was obtained from Clontech (Mountain View, CA) for gene expression studies.

Human adult heart tissue. Adult heart cDNA (Clontech, 637213 and 3 H Biomedical AB, SC6214) was used for measurement of gene expression. Tissue samples from adult left ventricles were obtained as surgical waste through an institutional review board at Columbia University.

Tissue bioreactor platform. The platform was assembled from two separate components: the wells for tissue culture, and an array of support structures with integrated elastomeric pillars for tissue attachment (1 mm in diameter, 6 mm axis-to-axis distance). Both components were fabricated out of polycarbonate using a computer numerical control (CNC) milling machine with mating features for stability and repeatable positioning (Extended Data Fig. 1a–c).

The pillars were formed by centrifugal casting of polydimethylsiloxane (PDMS, Dow Corning Sylgard 184) through, and extending from, the polycarbonate support structures. The supports were first inserted into Delrin (polyoxymethylene) moulds fabricated by CNC machining and polydimethylsiloxane (PDMS; 10:1 ratio of base:curing agent) was centrifugally cast at 400 relative centrifugal force for 5 min and cured in an oven at 60 °C for 1 h. The resulting component consisted of three pairs of pillars to support the formation of three tissues (Extended Data Fig. 1d). Pillars were 1 mm in diameter, 9 mm in length, and spaced 6 mm axis-to-axis.

The platform contained 12 wells for tissue culture that were patterned with exact 48-well-plate spacing, so that the platform corresponded to one quarter of the standard 48-well plate. Each well had a bottom portion measuring 10 mm × 4 mm × 4 mm where the cells in hydrogel were introduced, and a wider top portion measuring 10 mm × 7 mm × 4 mm for culture medium. A glass slide was bonded to the bottom of the platform to enable microscopic observation.

Electrical stimulation of the cell–hydrogel tissues was performed using carbon rods (Ladd Research Industries) as electrodes. The carbon rods were placed into slots machined on each side of the culture well, aligned in parallel and positioned perpendicular to the long axis of both the culture well and the tissue. The electrodes were connected to a cardiac stimulator (Grass s88x) by platinum wires (Ladd Research Industries). Electrical stimulation was generated by a spatially uniform, pulsatile electrical field (4.5 mV intensity, 2 ms in duration, monophasic square waveform) perpendicular to the long axis of the tissue. The parameter settings amplitude, duration, frequency and waveform were controlled by the Grass s88x cardiac stimulator.

Culture of cardiac tissues. Differentiated hiPS-CMs were combined with supporting human dermal fibroblasts (Lonza), cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% v/v fetal bovine serum, 100 U penicillin, and 0.1 mg/ml streptomycin, at a ratio of 75% hiPS-CMs and 25% fibroblasts. The cells were subsequently encapsulated in fibrin hydrogel by mixing 20 mg/ml human fibrinogen (Sigma), 100 U/ml human thrombin (Sigma-Aldrich) and the cell suspension at a 3:1:1 ratio. The hydrogel solution (200 μl containing 2 million cells) was dispensed into each well of the platform and allowed to polymerize at 37 °C for 30 min, so that the tissues readily formed around the pillars. Then, 800 μl of RPMI + B27 medium containing 0.2 mg/ml aprotinin (Sigma-Aldrich, A3428) were added into each well, with an additional 30 ml of RPMI + B27 medium containing 0.2 mg/ml aprotinin (Sigma-Aldrich, A3428) added to a 100-mm Petri dish (Corning, 430591) containing one platform (12 tissues). Subsequently, medium was changed every other day: 30 ml RPMI + B27 medium containing 0.2 mg/ml aprotinin (Sigma-Aldrich, A3428) for the first seven days, and then 30 ml RPMI + B27 medium (either days 7–28 or days 7–84).

The pillars were designed to subject the tissues to mechanical loading, mimicking that in native human myocardium. Hydrogel compaction caused passive tension in the tissues as they were stretched between the two pillars, inducing elongation and alignment. Synchronous contractions induced by electrical stimulation generated dynamic forces in the contracting tissues attached to the pillars that were forced to work against the load.

Electrical stimulation was initiated on day seven, using one of three training regimes (Fig. 1a): control (no electrical stimulation, 0 Hz), constant (constant frequency of 2 Hz), and intensity training (a ramped stimulation that increased the frequency, from 2 Hz on day 7 to 6 Hz on day 21, by 0.33 Hz per day; tissues were then stimulated at 2 Hz until day 28) (Extended Data Fig. 1e). Engineered tissues were randomly assigned to experimental groups. Tissues were cultured for a period of four weeks in 16 independent experiments, using three lines of iPSCs. Samples sizes for the main figures are shown in Supplementary Information, 'Main figure data sample sizes'.

Tissue properties were evaluated using real-time assessment of: amplitude and frequency of contractions, calcium handling, force generation, excitation threshold and maximum capture rate. End-point assays were performed to determine cell and tissue morphology (histologically), ultrastructure (by transmission electron microscopy), gene expression (using quantitative real-time PCR with reverse transcription (RT–PCR)) and the presence and distribution of cardiac proteins (immunohistochemistry).

Contractility analysis. Tissue contractility was measured by tracking the change in tissue area as a function of time. Live-cell, bright-field videos were acquired at rates of up to 150 frames per second using a Pike F-032b (Allied Vision Technologies) camera controlled with custom SPLASSH software[25]. Acquired video frames were inverted and an automated intensity threshold was used to identify cell location in the video frame. First, a baseline timepoint in the video corresponding to a relaxed tissue state was selected. Absolute differences in cell area from the baseline frame were then calculated to create a time course of cell area dynamics as a function of time. The resulting time courses were analysed using a native MATLAB automated peak finding algorithm to determine locations of maximum cell contraction indicated by the locations of local maxima in the timecourses. Beat period lengths were determined from the length of time between the pairs of local maxima. Beat frequencies were determined by inverting beat periods. Contraction amplitude relaxation times were measured from the length of time required for the tissue to relax from the peak contraction amplitude of the local maxima to the calculated relaxation amplitude (for example, the R90 time was the time elapsed between full contraction and 10% contraction).

Calcium handling. Tissues within culture platforms were loaded with Fluo-4 NW (50% v/v, Life Technologies) in RPMI + B27 medium containing 5 μM blebbistatin (Sigma) for 30 min at 37 °C as necessary to reduce movement artefacts. Videos were acquired at a rate of 150 frames per second using a Pike F-032 camera (Allied Vision Technologies) as described in 'Contractility analysis'. Videos were analysed in MATLAB using a custom script that calculated the temporal changes in calcium fluorescence intensity. Specifically, each frame was normalized to a baseline background region chosen by the user to give baseline-corrected changes in minimum and maximum fluorescence values for each frame. The temporal change in fluorescence intensity was presented as a calcium transient trace from which the measurements were obtained. In brief, the calcium transient 'timing' was determined as the peak-to-peak values of two successive beats as defined by the peak maxima. Calcium transient 'amplitude' was determined by numerically integrating the area below the peak maxima relative to the baseline. Calcium transient traces were analysed during 5-mM-caffeine stimulation of tissues previously treated with either 1 mM verapamil (Sigma-Aldrich) or 1 μM thapsigargin (Sigma-Aldrich). Caffeine responses were quantified by comparing this calcium transient amplitude before and after the addition of 5 mM caffeine (Sigma-Aldrich).

Conduction velocity. A surrogate of conduction velocity was assessed by calcium propagation within the entire tissues that were pre-treated with 5 μM blebbistatin (Sigma-Aldrich) to uncouple true Ca^{2+}-dependent fluorescent motion from the fluorescent signals caused by motion artefacts. The conduction velocity was calculated by selecting two sections of the tissue within the region of calcium transient propagation and dividing the distance between the centres of these regions by the difference between their peak maxima.

Direct measurements of force. The force generation was measured directly, using an organ bath with high-sensitivity force transducers. Cardiac tissues and FCT strips were transferred to a commercial organ bath system (DMT Myograph) containing oxygenated modified Tyrode's solution (129 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$, 30 mM glucose, 25 mM HEPES, pH 7.4) supplemented with 2% B27 and maintained at a constant temperature of 37 °C without electrical stimulation. All measurements were done using LabChart software (ADInstruments). The tissues were allowed to equilibrate for 15 min and any spontaneous beating measurements were recorded. The tissues were then allowed to equilibrate for another 15 min under electrical stimulation (2 Hz, 5 ms, 80–100 mA, rectangular pulses) in order to preload the tissues by manual stepwise adjustment of the tissue length to that of the maximal force generated, which assumes the optimal sarcomere length is thereby attained.

Twitch tension was measured by increasing the organ bath $[Ca^{2+}]$ from 0.2 to 2.8 mmol/l. Specifically, the extracellular calcium concentration was changed by changing the concentration of $CaCl_2$ used in the Tyrode's solution. The tissues were subjected to electrical stimulation for 3 min, and an average of 10 contractions were measured. The stimulation was then discontinued for 10, 20 or 30 s, and the tissues were allowed to recover for 2 min. Post-rest potentiation measurements were subsequently obtained by analysing the change in twitch tension from the first beat upon re-initiation of electrical stimulation.

Contractility and twitch parameters were further investigated in response to the increasing electrical stimulation frequency within the organ bath as previously described[25]. Twitch forces were calculated as the average of the difference between cyclic peak maximum and minimum force and normalized to the cross-sectional area (obtained by measurement of tissue at the centre after force measurements). The force–frequency relationship was measured by increasing the electrical stimulation frequency from 1 Hz to 6 Hz in 1-Hz increments. The tissues were subjected to each stimulation frequency for 30 s before increasing to the next stimulation frequency. The force data were measured at frequencies of 1–6 Hz (in 1 Hz increments) for all experimental groups (static, constant, early-stage and late-stage intensity-trained tissues, and human fetal strips) and all iPSC lines.

Continuous recordings of force and calcium as a function of frequency. Continuous videos were recorded at a rate of 100 frames per second with a Zyla 4.2 sCMOS camera (Andor) to determine calcium transients and tissue displacement. The stimulation frequency was increased from 1 Hz to 6 Hz in 1-Hz increments every 20 s (that is, every 2,000 frames). The calcium transients were analysed using custom MATLAB software as described above for measurements of calcium traces, and normalized to the baseline at each frequency as $(F–F_0)/F_0$. Tissue displacement was measured using the Spottracker module in ImageJ. The areas within the tissue were manually selected at baseline and tracked frame-to-frame to measure changes in the pixel displacement over time. Calcium dye loading was performed as previously described, but without the use of blebbistatin to block contractile motion. This enabled measurements of both calcium transient intensity and displacement during calcium imaging.

Immunofluorescent staining. For morphological analysis, tissues and FCTs were fixed by using gradually increasing concentrations of paraformaldehyde (1–4%, in 1% increments, 1 h each). Whole tissues were paraffin-embedded and cut into 5-μm-thick sections. The sectioned tissues were quenched in 0.5 M NH_4Cl for 30 min, permeabilized with 0.2% Triton X-100 in PBS for 15 min and then incubated in blocking solution (1% bovine serum albumin (BSA), 2% goat serum in PBS) for 2 h. The following primary antibodies were incubated for 2 h in 1% BSA: anti-sarcomeric α-actinin (1:200; Abcam ab9465), anti-cardiac troponin T (cTnT, 1:100; Thermo Scientific MS-295-P1), anti-ryanodine receptor 2 (RYR2, 1:100; Abcam ab2827), anti-Ca_V1.2 (1:200; Abcam ab58552), anti-BIN1 (1:100; Abcam ab137459), anti-mitochondria (1:50; Abcam ab3298) and anti-OXPHOS (1:100; Acris MS601-720). Actin was deteced with Alexa Fluor 350–phalloidin (Thermo Fisher A22281).

Tissues were washed three times for 5 min in 0.2% Triton X-100 and incubated with the corresponding secondary antibodies for 2 h: anti-mouse IgG–Alexa Fluor 488 (1:400; Invitrogen A21202), anti-rabbit IgG–Alexa Fluor 568 (1:400, Invitrogen, 81-6114) and anti-mouse IgG–Alexa Fluor 635 (1:400, Invitrogen, A31574). The tissues were washed and subsequently incubated with NucBlue (Molecular Probes, R37606) for nucleus counterstaining. The immunostained tissues were visualized using a confocal microscope (Olympus Fluoview FV1000).

For T-tubule immunostaining, tissues were incubated with WGA–Alexa Fluor 488 (Life Technologies, W11261) or di-8-ANEPPS (Life Technologies, D-3167) for

20 min before permeabilization and subsequent staining with additional antibodies as described above.

Transmission electron microscopy. Tissues, FCTs, and adult heart tissue were fixed with 2.5% glutaraldehyde in 0.1 M Sorenson's buffer (pH 7.2) for 1 h and sent to the Electron Microscopy and Histology (EM&H Core) Facility at Weill Cornell Medical College for subsequent sample preparation, imaging and data interpretation in a blinded fashion. Samples were post-fixed for an additional hour with 1% OsO_4 in Sorenson's buffer. After dehydration, the samples were embedded, sectioned, stained with toluidine blue and examined under a JEM-1400 electron microscope.

Fraction of cells containing sarcomeres. Sectioned tissues were immunoflourescently labelled with sarcomeric α-actinin and DAPI. Using the cell counter plugin in ImageJ, the DAPI-positive cells were marked and counted. Subsequently, all DAPI-positive cells that stained positive for α-actinin were counted, and the percentage of cells containing sarcomeres of DAPI-positive cells was calculated.

Sarcomere length. Sarcomere length was determined in dissociated cells that were replated as a monolayer and stained with sarcomeric α-actinin by measuring the distance between intensity peaks along the long axis of designated cell areas containing clear striations[9]. A minimum of three sarcomere lengths per cell were obtained in large numbers of cells from $n > 6$ biological replicates.

Change in tissue area. The change in the projected tissue area (per cent change between the contracted and relaxed state) was experimentally determined in bright light by analysing the change tissues paced at 1 Hz and twice the excitation threshold, by custom-designed MATLAB code that used video edge-detection based on the contrast between the darker tissue and the lighter surrounding area. For each group and time point, the change in area was normalized to the change in area measured at day 6, immediately before the application of electrical stimulation.

Cell morphology. Cells were enzymatically digested using serial digestions of collagenase type 1 and 2 (Worthington), and plated onto eight-well chamber slides (Laboratory-Tek, Sigma-Aldrich). The cells were allowed to attach for 72 h and imaged using phase-contrast microscopy. Cell area was quantified from the images using the '%Area' function in ImageJ after thresholding of the cells in each image. Cell elongation ratio was calculated from these images using the 'Roundness' function in ImageJ, in which the aspect ratio was defined as (1 − Roundness), with 0 corresponding to a circle and 1 corresponding to a completely elongated object[30].

RT–PCR. Total RNA was purified from tissues according to the manufacturer's instructions using TRIzol (Life Technologies). For measurements of adult and FCT *NPPA* and *NPPB* expression, commercial tissues were used: adult heart, ages: 30–39, pooled from three male hearts (TaKaRa/Clontech Human RNA Master Panel II, 636643, lot no. 1208462 A); fetal hearts, GW: 21–37, pooled from 32 male and female fetal hearts (Clontech Human Fetal Heart Poly-A$^+$ RNA, 636156, lot no. 7110214; synthesized with oligo-dT20 and SSIII kit). Reverse transcription was performed using Ready-To-Go You-Prime First-Strand Beads (GE Healthcare, 27-9264-01) following the manufacturer's instructions. Gene expression was quantified by real-time PCR using SYBR Green primers (Life Technologies) in an Applied Biosystems Step One Plus. Data analysis was carried out using the $\log_2$-fold change normalized to late-stage week 1 tissue gene expression shown in Fig. 1b, Extended Data Fig. 2a. Data analysis was carried out using the fold change normalized to glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene expression in Extended Data Fig. 2b, 9c. Primers used are listed in Supplementary Information, 'Primer list'.

Oxygen consumption rate and extracellular acidification rate. Engineered cardiac tissues were dissociated into single cells after four weeks of cultivation using activated papain solution containing 20 U/ml papain (~15 min at 37 °C with gentle tapping), as described for electrophysiological recordings. The enzyme reaction was terminated by adding 10% FBS in DMEM/F-12 culture medium. The dissociated hiPS-CMs were plated into XF96 Culture Plates (Seahorse Bioscience) coated with Matrigel (Corning, 354230) and cultured for three days. Subsequently, the plated hiPS-CMs were assayed in real-time using an XF-96 Extracellular Flux Analyzer (Seahorse Biosciences) for oxygen consumption rate (OCR) and extracellular acidification rate (ECAR) per the manufacturer's protocols.

Isoproterenol response. Isoproterenol was diluted in standard medium (RPMI + B27) to a concentration of 1 μM. Tissues were placed in the organ bath (as previously described) and equilibrated for 10 min. Videos were captured, force measurements were recorded before and after addition of the drug, and the change in the generated force was determined.

Patch-clamp electrophysiology. Engineered cardiac tissues were dissociated into single cells for whole-cell patch-clamp recordings at four weeks using activated papain solution containing 20 U/ml papain from *Caripa papaya* (Sigma-Aldrich 76220), 1.1 mM EDTA, 67 μM 2-mercaptoethanol (Sigma-Aldrich M3148), 5.5 mM L-Cysteine-HCl (Sigma-Aldrich C7880) in 1 × EBSS (Thermo Scientific/ Gibco 24010-043). This optimized protocol enabled healthy patchable single cardiomyocytes to be obtained without spontaneous beating from early-stage

intensity-trained cardiac tissues. The same dissociation protocol using papain was also used for the other cardiac tissue samples. The tissues were incubated for ~15 min at 37 °C with gentle tapping, and the enzyme reaction was terminated by adding FBS (10%) in DMEM/F-12 culture medium.

Whole-cell patch-clamp recordings of dissociated iPSC-CMs were conducted using a patch-clamp amplifier (MultiClamp 700B, Molecular Devices) and an inverted microscope equipped with differential interface optics (Nikon, Ti-U). Glass pipettes were prepared from borosilicate glass pipettes (Sutter Instrument BF150-110-10) and a micropipette puller (Sutter Instrument, Model P-97).

Current-clamp recording for action potential measurements was conducted in normal Tyrode's solution containing 140 mM NaCl, 5.4 mM KCl, 1 mM $MgCl_2$, 10 mM glucose, 1.8 mM $CaCl_2$ and 10 mM HEPES (pH 7.4 with NaOH at 25 °C) using the pipette solution 120 mM K D-gluconate, 25 mM KCl, 4 mM MgATP, 2 mM NaGTP, 4 mM Na_2-phospho-creatine, 10 mM EGTA, 1 mM $CaCl_2$ and 10 mM HEPES (pH 7.4 with KCl at 25 °C). Action potentials were stimulated (5 ms, 0.3 nA) in a current clamp mode at 37 °C (0.2 Hz), recorded and analysed using Clampfit 10.4 (Axon Instruments).

Voltage-clamp measurements for I_{K1} current recording were conducted using an extracellular solution containing 160 mM NMDG, 5.4 mM KCl, 2 mM $MgCl_2$, 10 mM glucose, 10 μM nisoldipine, 1 μM E-4031 and 10 mM HEPES (pH 7.2 with HCl at 25 °C) and a pipette solution 150 mM K-gluconate, 5 mM EGTA, 1 mM Mg–ATP and 10 mM HEPES (pH 7.2 with KOH at 25 °C). The following pulse protocols were used: 2-s voltage clamp applied from -130 to $+10$ mV (holding at -40 mV, 0.1 Hz, 2-s voltage pulse). The I_{K1} reversal potential (Ba^{2+}-sensitive current) had a negative slope conductance consistent with inward rectification, as previously described[30]. The current–voltage plot was analysed before and after the addition of 0.5 mM $BaCl_2$ for 2 min.

Dose–response curves. Drugs were diluted in standard medium (RPMI + B27). Successively higher doses of each drug were administered at concentrations of 10^{-11} M to 10^{-5} M, in decigram increments. Videos were captured ≥5 min after each dose was administered, and processed using custom image processing software as described above. For chronotropic drugs, bright-field videos were taken at each drug concentration so that measurements of the beat frequency could be determined as a function of drug concentration. For ionotropic drugs, tissues were placed in the organ bath and force measurements were recorded as previously described at each drug concentration so that the measurements of the change in force generated could be determined as a function of drug concentration. Dose–response curves for these parameters could be constructed by using MATLAB to fit the Hill equation for sigmoid curves to the data, to determine the corresponding EC_{50} value.

Paced isoproterenol response. Cardiac tissues were loaded with calcium dye as described above. Tissues were transferred to standard medium (RPMI + B27), paced at 1 Hz for 30 min to equilibrate, and the baseline video recordings were then obtained. Successively higher doses of isoproterenol were administered directly to the standard medium at concentrations of 0.01, 100 and 1,000,000 nM. Videos of tissues were captured ≥10 min after each dose was administered, and processed using custom-designed image analysis software as described above.

Cardiac hypertrophy model. Cardiac tissues were exposed to drugs known to induce pathological hypertrophy (angiotensin II, endothelin-1, isoproterenol) on day 6 following tissue formation. The first time point (1 week) was taken after 24 h of incubation with or without the drug. The majority of the HCM data shown are from pathologically induced HCM via endothelin-1 addition (unless data from all three drugs are shown), since the results were comparable amongst the three HCM-inducing agents.

Statistics and reproducibility. Data are shown as mean ± 95% CI. Differences between experimental groups were analysed by one-way or two-way ANOVA. Post hoc pairwise analysis was done using Tukey's HSD test. Electrophysiological data were analysed by one-way ANOVA Barlette's test with multiple comparisons. P values < 0.05 were considered significant for all statistical tests. The reproducibility of the data are demonstrated by the number of independent biological samples. The number of independent experiments performed for each dataset reported in the main figures is detailed in Supplementary Information, 'Main figure data sample sizes'. Details of the sample sizes used in the Extended Data are included in the respective figure legends.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The study used a combination of commercial and open-source software packages, which are specified in the Methods, and custom-designed software that will be made available to interested investigators upon reasonable request.

Data availability. Source data for quantitative data shown in all figure panels are available without restrictions and can be accessed at https://doi.org/10.6084/m9.figshare.5765559. The detailed experimental protocol is available from Protocol Exchange[31].

30. Ma, J. et al. High purity human-induced pluripotent stem cell-derived cardiomyocytes: electrophysiological properties of action potentials and ionic currents. *Am. J. Physiol. Heart Circ. Physiol.* **301**, H2006–H2017 (2011).
31. Morikawa, K., Song, L., Ronaldson-Bouchard, K., Vunjak-Novakovic, G. & Yazawa, M. Electrophysiological recordings of cardiomyocytes isolated from engineered human cardiac tissues derived from pluripotent stem cells. *Protoc. Exch.* https://doi.org/10.1038/protex.2018.030 (2018).
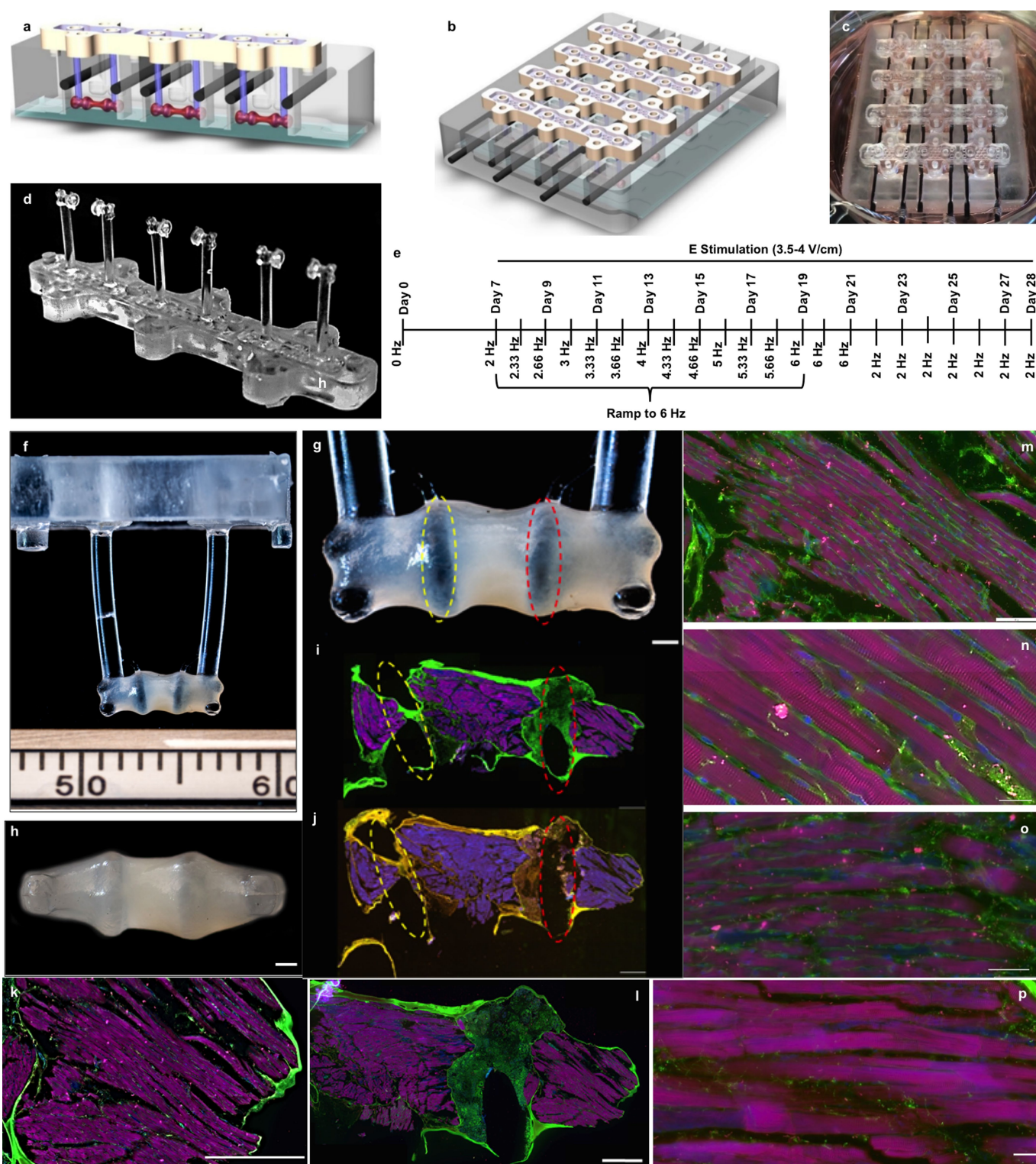
Extended Data Fig. 1 | Experimental design and overall appearance of cardiac tissues. a, Schematic of the pillars (purple) placed via interlocking mating components between the bioreactor wells (grey) and pillar lid (yellow) with tissues (pink) formed around the pillars, and electrodes (black) placed perpendicular to the cardiac tissues. A glass slide (blue) is epoxied to the bottom of the bioreactor to enable image acquisition. **b**, A schematic of the assembled bioreactor. **c**, Photographs of the cardiac tissues cultured within the bioreactor. **d**, The tissue pillar. **e**, Increase in the electrical stimulation frequency throughout the intensity training regime. **f–h**, Photographs of the tissues attached to pillars at the end of four-week cultivation: side view (**f**, **g**), bottom view (**h**). Scale bars, 500 μm. **i**, **j**, Immunofluorescence in serial sections of the early-stage intensity-trained tissue. The dotted yellow and red lines in **g**, **i**, **j** indicate corresponding pillar placement within the tissue. Scale bars, 500 μm. **k–p**, Immunofluorescence in serial sections of the early-stage intensity-trained tissue in **i**. WGA, green; α-actinin, pink; nuclei, blue. Scale bars, **i–l**, 500 μm; **m**, 100 μm; **n**, 20 μm; **o**, **p**, 50 μm. Images were selected to include landmark features that facilitate localization and comparisons. Similar results were obtained from three independent experiments.
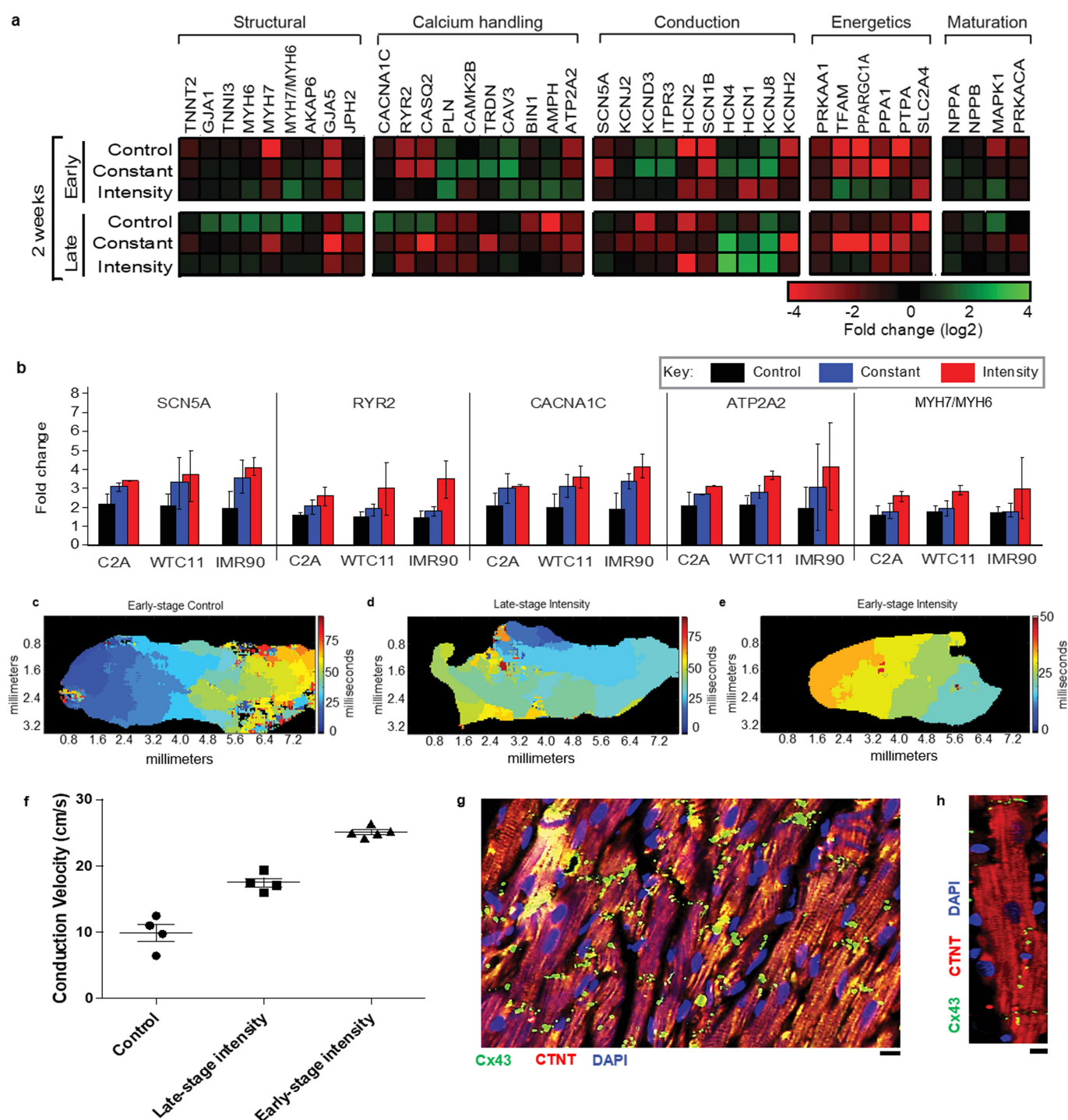
Extended Data Fig. 2 | Enhanced gene expression and conduction in intensity-trained cardiac tissues over time. a, Quantitative gene expression in FCTs and C2A iPSC cardiac tissues after two weeks of culture, as determined by RT–PCR; shown as fold change relative to late-stage tissues at the start of stimulation. **b,** Quantitative gene expression in early-stage cardiac tissues, normalized to *GAPDH*, from three different iPSC lines as determined by RT–PCR after four weeks of culture. $n = 12$ biologically independent samples per group; mean ± 95% CI; no significant difference between the cell lines by two-way ANOVA. **c–f,** Representative conduction velocity activation maps for early-stage control (**c**), late-stage intensity-trained (**d**) and early-stage intensity-trained (**e**) cardiac tissues, and surrogate of conduction velocity in early-stage and late-stage C2A iPSC cardiac tissues after four weeks of culture, assessed by calcium propagation (**f**). Mean ± s.e.m., $n = 4$–5 biologically independent samples per group. **g, h,** Representative immunofluorescence of gap junction (connexin-43 (Cx43), green) expression in early-stage intensity-trained iPSC cardiac tissue after four weeks of culture, at low (**g**; scale bar, 10 μm) and high magnification (**h**; scale bar, 5 μm). cTnT, red; nuclei (DAPI), blue. Similar results were obtained from four independent experiments.
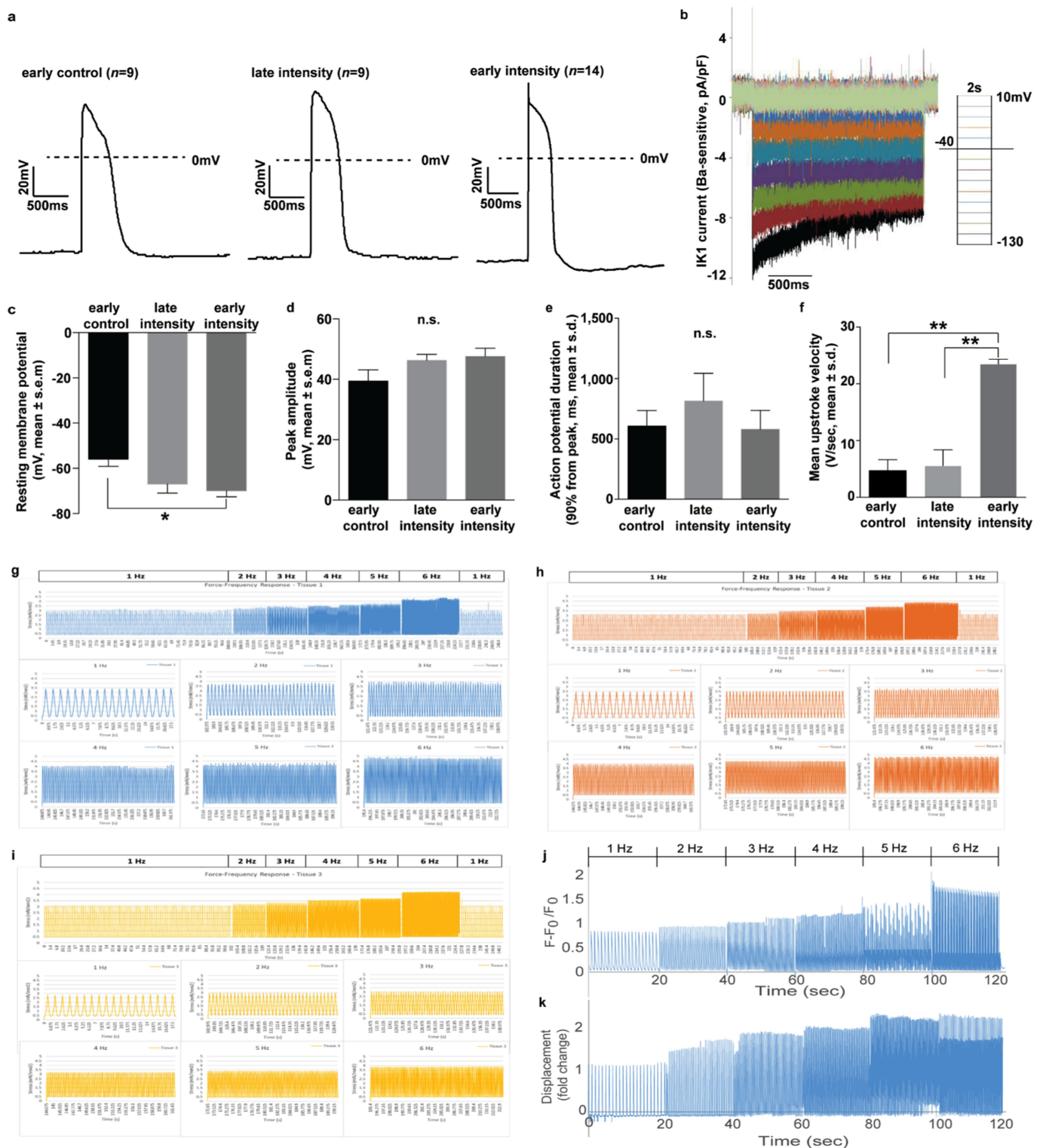
Extended Data Fig. 3 | Electrophysiological characterization of human engineered cardiomyocytes. a, Representative traces of action potentials in early-stage control ($n = 9$), late-stage intensity-trained ($n = 9$) and early-stage intensity-trained ($n = 14$) groups. n is the number of biologically independent samples obtained during two independent experiments. **b**, Representative traces of I_{K1} current for the early-stage intensity-trained group using voltage-clamp mode. **c–f**, Electrophysiology data after four weeks of culture showing the resting membrane potential (**c**), peak amplitude (**d**), duration of action potential (**e**) and upstroke velocity (**f**) obtained in two independent experiments, resulting in biologically

independent data from early-stage control ($n = 9$), late-stage intensity-trained ($n = 9$) and early-stage intensity-trained ($n = 14$) groups. **$P < 0.01$, *$P < 0.05$ using one-way ANOVA Bartlett's test with multiple comparison. n.s., not significant. **g–i**, Representative continuous organ bath force recordings under electrical pacing from 1–6 Hz from three biologically independent early-stage intensity trained tissues (C2A cells) from one experiment. **j**, **k**, Representative continuous recordings from an early-stage intensity-trained tissue (C2A cells) under electrical pacing from 1–6 Hz of calcium (**j**) and surrogate force (**k**) as determined by tissue displacement, normalized to 1 Hz.

a FCT

Key: —— FCT —— Control —— Constant —— Intensity

b Early-stage

Control Constant Intensity

2 wks

3 wks

4 wks

c Late-stage

Control Constant Intensity

2 wks

3 wks

4 wks

d

Key: FCT Control Constant Intensity

e

Group	MLC2v+	MLC2a+	MLC2v+/MLC2a+
Early-stage Control	51.7	18.8	2.75
Early-stage Intensity	65.0	5.5	11.8
Late-stage Control	37.9	15.9	2.39
Late-stage Intensity	41.1	12.7	3.23

f Unstained 0.00% | CTNT 70.5%

g Unstained 0.17% | Vimentin 24.5% | vWF 4.95%

h α-actinin, CTNT, DAPI **i** α-actinin, CTNT, DAPI **j** α-actinin, CTNT, DAPI **k**

Extended Data Fig. 4 | Enhanced maturation and synchronicity of cardiac tissues in response to training regime as a function of time. **a–c,** Representative contraction profiles of FCT (**a**), early-stage (**b**) and late-stage cardiac tissues (**c**) over time (C2A cell line). **d,** Frequency of contractions in cardiac tissues over four weeks of culture. $n = 35$ biologically independent samples over 16 independent experiments; mean $\pm$ 95% CI, *$P < 0.05$ compared to control group by two-way ANOVA with Tukey's HSD test. Early-stage intensity-trained tissue shows significant differences versus other training regimes by two-way ANOVA with Tukey's HSD test. **e,** Characterization of cardiac cell population within cardiac tissues (C2A line) after four weeks of culture by fluorescence-activated cell sorting (FACS) analysis. **f, g,** Characterization of cells isolated from early-stage intensity-trained cardiac tissues (C2A line) by FACS analysis after four weeks of culture: cardiac cells (**f**; cTnT), and supporting fibroblast cells and endothelial cells (**g**; vimentin and von Willebrand Factor(vWF), respectively). **h–j,** Representative immunofluorescence of whole tissues showing the enhanced cardiac ultrastructure (α-actinin, green; cTnT, red; nuclei, blue) in early-stage cardiac tissues from the C2A line (**h**), WTC11 cell line (**i**), and IMR90 cell line (**j**) after four weeks of culture. Scale bars, 5 μm; experiment repeated independently 14 times with similar results. **k,** Representative immunofluorescence showing the cell population in a histological section from early-stage cardiac tissue (C2A line) after four weeks of culture. cTnT, green; vimentin, red; nuclei, blue. Scale bar, 50 μm; experiment repeated independently two times with similar results.
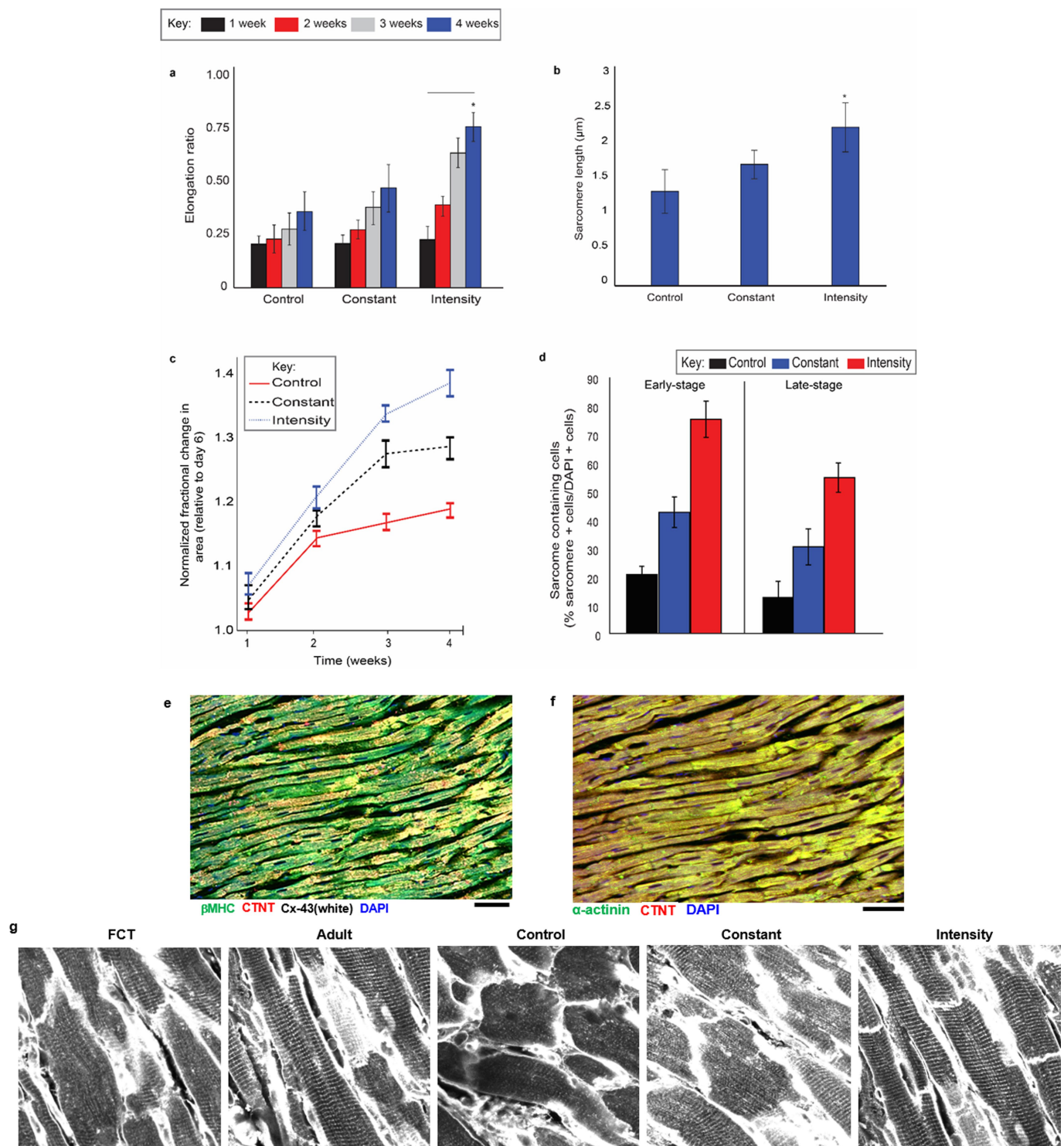
Extended Data Fig. 5 | Physiological hypertrophy within cardiac tissues enhances contractility. a–c, Physiological hypertrophy of cardiomyocytes cultured in the electromechanically conditioned cardiac tissue format increases as a function of time and training regime beyond FCT levels, as shown by cell elongation ratio (**a**) and sarcomere length (**b**). $n = 326$ biological replicates from 15 independent samples in one experiment. **c,** This enables the change in area while being electrically paced at 1 Hz, an indirect measure of fractional shortening, to similarly increase beyond FCT levels as a function of time and training regime. Data represent the ratio of the change in area for a given time point and the change in area at day 6. $n = 6$ biologically independent samples per group; mean $\pm$ 95% CI; *$P < 0.05$ compared to FCT group at week four by ANOVA with Tukey's HSD test; line above graph indicates $P < 0.05$ compared to other training

regimes by two-way ANOVA with Tukey's HSD test. **d,** The enhanced cardiac ultrastructure in intensity-trained early-stage cardiac tissues is documented by the quantification of sarcomere distribution in cardiac tissues. $n = 12$ biologically independent samples per group, mean $\pm$ 95% CI. **e, f,** Representative immunofluorescence of gap junction (connexin-43 (Cx43), white) in early-stage iPSC cardiac tissue (β-myosin heavy chain (β-MHC), green; cTnT, red; nuclei (DAPI), blue) (**e**) and cardiac ultrastructure in early-stage iPSC cardiac tissue (α-actinin, green; cTnT, red; nuclei (DAPI), blue) (**f**) after four weeks of culture. Scale bar, 50 μm; experiment repeated independently three times with similar results. **g,** α-Actinin immunofluorescence (white) in cardiac tissues after four weeks of culture. Scale bar, 10 μm; experiment repeated independently two times with similar results.
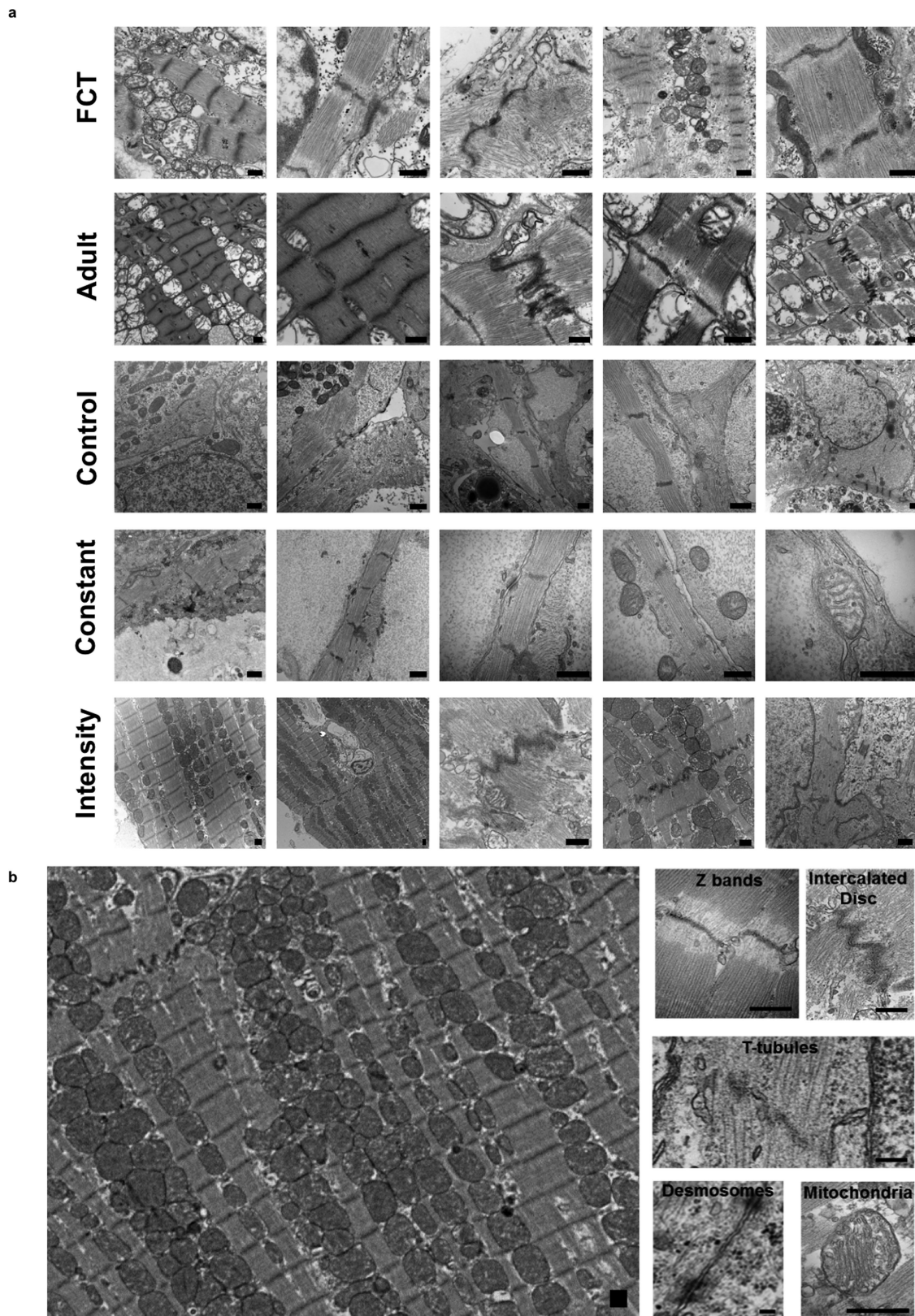
Extended Data Fig. 6 | Enhanced ultrastructural properties of cardiac tissues following intensity training. a, Representative transmission electron microscopy images for FCTs, adult cardiac tissue, and early-stage cardiac tissues (C2A line) using different electromechanical conditioning protocols, after four weeks of culture. Scale bar, 500 nm. **b**, TEM images of intensity-trained early-stage cardiac tissues (C2A line) after four weeks of culture showing details of various ultrastructural elements, Scale bar, 500 nm. Similar results to those in **a** and **b** were obtained independently with the following cells or treatments: FCT ($n = 8$), adult ($n = 2$), control ($n = 3$), constant ($n = 3$), intensity-trained ($n = 4$).
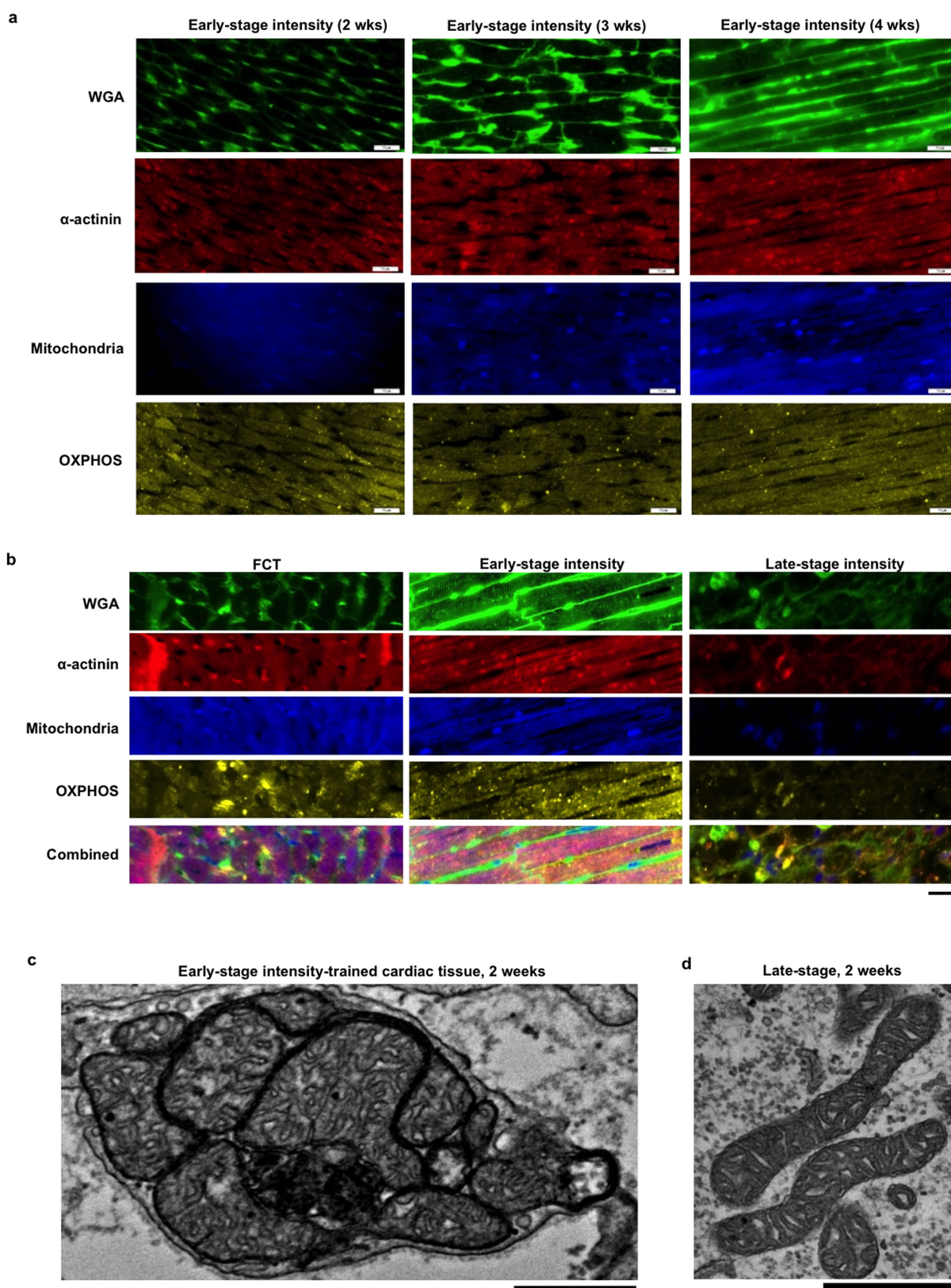
Extended Data Fig. 7 | Intensity training of cardiac tissues derived from early-stage hiPS-CMs is required to enhance mitochondrial development. a, Representative immunofluorescence showing ultrastructural proteins WGA (green), α-actinin (red), mitochondria (blue) and oxidative phosphorylation (yellow) for early-stage cardiac tissues (C2A cell line) at different culture times during exposure to the intensity-training electromechanical-conditioning regime. Scale bar, 20 μm. **b**, Representative immunofluorescence showing ultrastructural proteins WGA (green), α-actinin (red), mitochondria (blue) and oxidative phosphorylation (yellow) in cardiac tissues cultured with intensity training for four weeks from early-stage hiPS-CMs (C2A cell line), late-stage hiPS-CMs (C2A cell line) and GW19 FCT. Scale bar, 20 μm. Similar results to those in **a** and **b** were obtained independently from the following experiments: FCT ($n = 5$), early-stage intensity-trained ($n = 3$), late-stage intensity-trained ($n = 3$). **c**, **d**, Representative TEM images for early-stage (**c**) and late-stage cardiac tissues (**d**) (C2A cell line) after two weeks of exposure to the intensity-training electromechanical-conditioning regime. Scale bar, 1 μm; experiment not repeated independently.
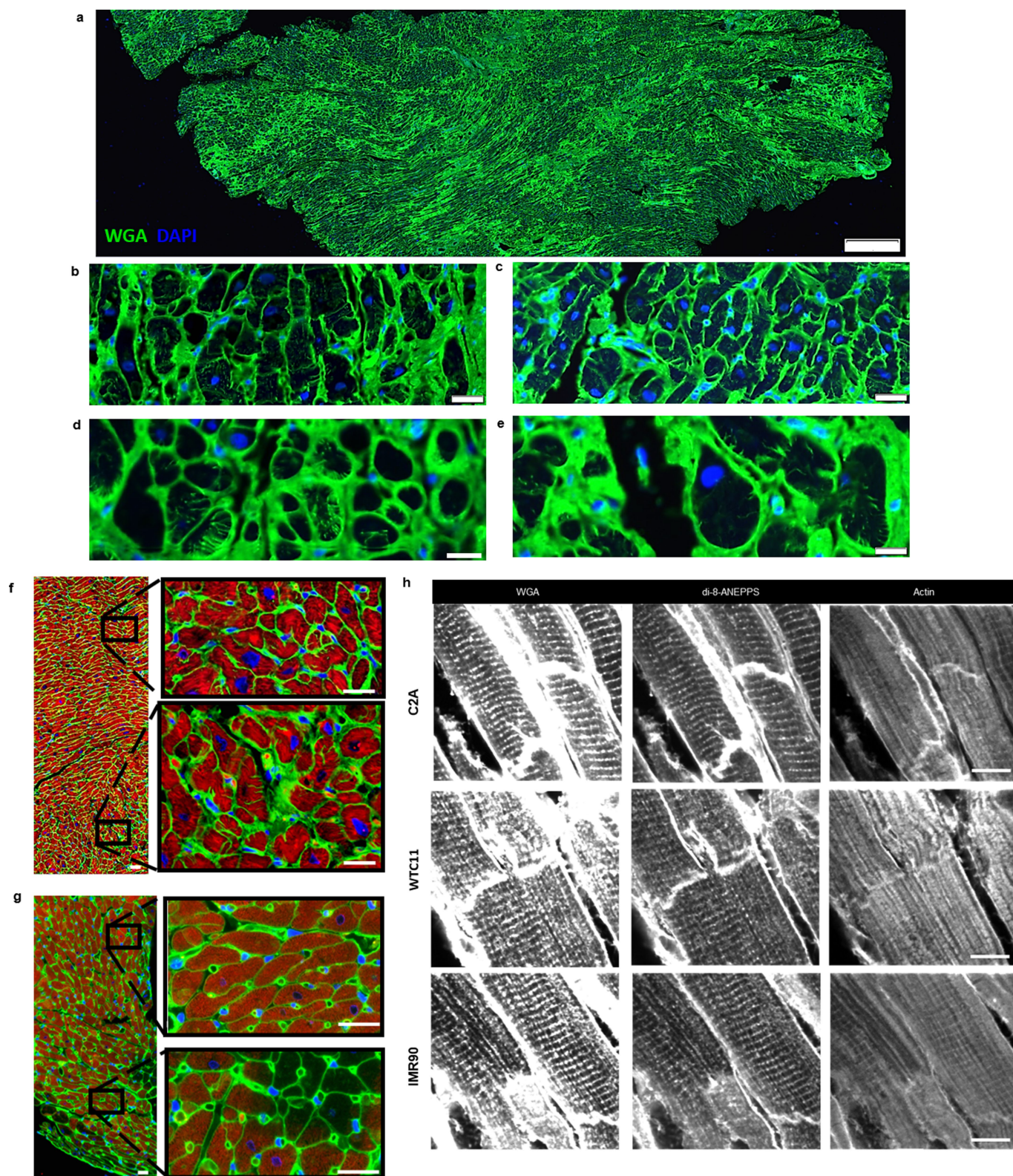
Extended Data Fig. 8 | Formation of T-tubules in early-stage intensity-trained cardiac tissues. a–e, Axial tissue cross-sections from intensity-trained cardiac tissues (C2A line) after four weeks of culture showing T-tubules (WGA, green) and nuclei (DAPI, blue) at low magnification (**a**; scale bar, 100 μm), medium magnification (**b** and **c**; scale bar, 10 μm) and high magnification (**d** and **e**; scale bar, 5 μm). **f, g,** Axial tissue cross-sections showing T-tubules (WGA, green), actin (red) and DAPI (blue) in intensity-trained cardiac tissues (C2A line) after four weeks of culture (**f**)

and GW19 FCT (**g**). Scale bar, 10 μm. **h,** Immunofluorescence of paraffin-embedded and sectioned cardiac tissues from three different iPSC cell lines (C2A, WTC11, IMR90) after four weeks of intensity training showing the formation of T-tubules (confirmed by both WGA staining and di-8-ANEPPS staining), and striated ultrastructure (actin). Scale bar, 10 μm. Similar results to those in **a–h** were obtained in a minimum of four independent experiments.
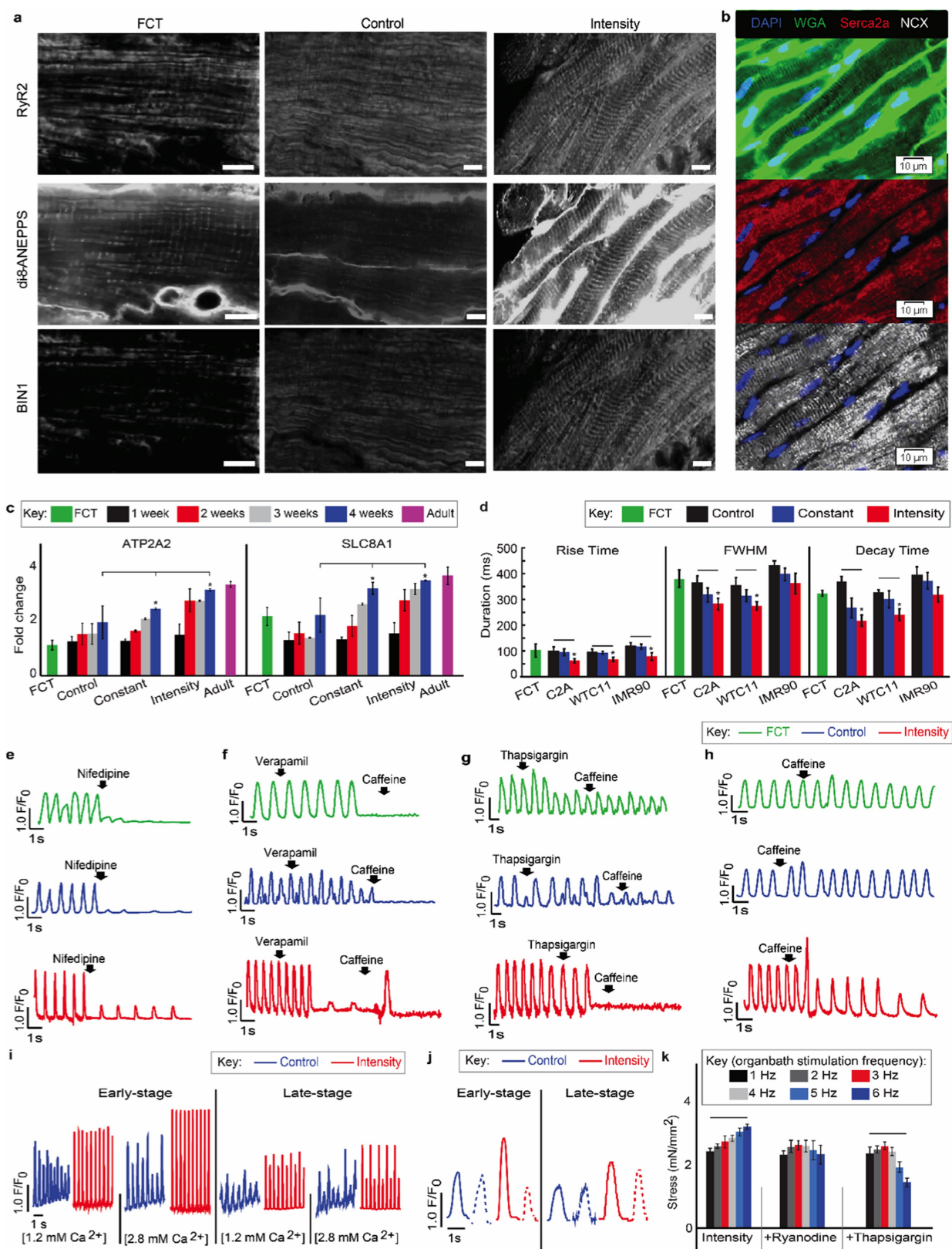
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Intensity training upregulates cardiac maturation in early-stage tissues through enhanced calcium handling. **a**, **b**, Intensity training promotes T-tubule formation in early-stage hiPS-CM tissues, as demonstrated by immunofluorescence of ryanodine 2 receptor (RYR2, green), bridging integrator 1 (BIN1, blue) and T-tubule staining (di-8-ANEPPs, red). Scale bar, 10 μm. **c**, Expression of *ATP2A2* and *SLC8A1* genes, which are responsible for maintaining proper calcium homeostasis, in early-stage tissues as determined by RT–PCR and normalized to *GAPDH* over four weeks of culture with the designated stimulation regime. Independent biological replicates per group: FCT, $n = 8$; control, $n = 6$; constant, $n = 6$; intensity-trained, $n = 14$; adult, $n = 1$. Mean ± 95% CI, *$P < 0.05$ versus FCT group at week four by ANOVA with Tukey's HSD test. Line over graph indicates $P < 0.05$ compared to other training regimes by two-way ANOVA with Tukey's HSD test. **d**, Relaxation times in early-stage tissues as characterized by the full-width half-maximum (FWHM) values and the decay time (90% of the time from the maximal peak of the calcium transient). Independent biological replicates per group: FCT, $n = 8$; C2A, $n = 12$; WTC11, $n = 6$; IMR90,

$n = 6$. Mean ± 95% CI; *$P < 0.05$ versus FCT group by ANOVA with Tukey's HSD. Line over graph indicates $P < 0.05$ between cell lines by two-way ANOVA. **e**, Representative calcium traces of early-stage tissues treated with 1 μM nifedipine. **f**, **g**, Representative traces of calcium release after stimulation with 5 mM caffeine in early-stage tissues and FCTs treated with 1 mM verapamil (**f**) or 2 μM thapsigargin (**g**). **h**, Representative traces of calcium release after stimulation with 5 mM caffeine for early-stage tissues and FCTs. **i**, Calcium spikes detected by fluorescent calcium dyes in early and late-stage tissues (C2A line) after four weeks of culture at two calcium concentrations. **j**, Intensity-trained early-stage but not late-stage tissues (C2A line) after four weeks of culture respond to ryanodine (1 μmol l^{-1}). **k**, The force–frequency relationship of early-stage intensity-trained cardiac tissues (C2A line) after four weeks of culture, treated with the RYR2 blocker ryanodine (1 μM) or the SERCA2a blocker thapsigargin (1 μM). Directly measured force data; $n = 13$ biologically independent samples for intensity group and $n = 3$ biologically independent samples for other groups. Mean ± 95% CI, line over graph indicates $P < 0.05$ by two-way ANOVA.
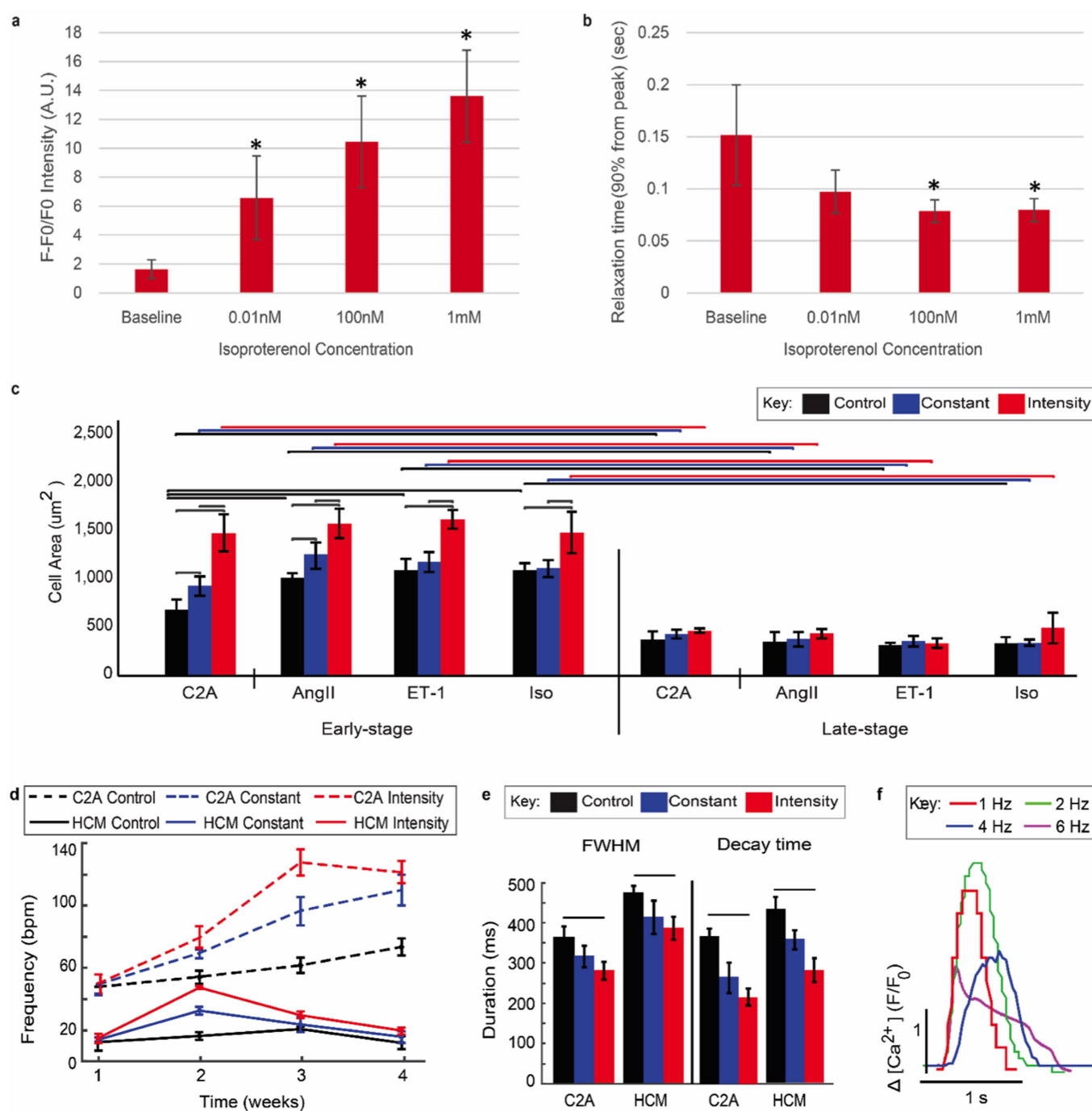
Extended Data Fig. 10 | Intensity training in early-stage tissues enables physiologically relevant drug responses and the development of a pathological hypertrophy disease model. a, **b**, Calcium intensity measurements (**a**) and relaxation time obtained by measuring the time from the peak to 90% of the relaxation (R90) during electrical pacing (**b**) at 1 Hz in early-stage intensity-trained tissues (C2A line) after four weeks of culture with increasing doses of isoproterenol. $n = 20$ biological replicates from six independent experiments. Mean ± 95% CI; *$P < 0.05$ versus baseline response by ANOVA with Tukey's HSD test. **c**, Cell area over four weeks of culture for the designated stimulation regime. $n = 10$ biological replicates from five independent experiments. Mean ± 95% CI; line above graph indicates $P < 0.05$ compared to other training regimes by two-way

ANOVA with Tukey's HSD test. **d**, Frequency of contractions in healthy (C2A) and hypertrophic (HCM) heart tissues over four weeks of culture. $n = 12$ independent biological samples from five independent experiments. Mean ± 95% CI. **e**, Relaxation times in early-stage tissues (C2A line) and early-stage hypertrophy tissues (HCM) as characterized by FWHM values and the decay time (90% of the time from the maximal peak of the calcium transient). $n = 20$ biological replicates from four independent experiments. Mean ± 95% CI; line above graphs indicate $P < 0.05$ compared to other training regimes by two-way ANOVA with Tukey's HSD test. **f**, Early-stage intensity-trained hypertrophy tissues exhibit impaired FDAR, as shown for each stimulation frequency by individual traces of calcium peaks.

LETTER

Distributed hepatocytes expressing telomerase repopulate the liver in homeostasis and injury

Shengda Lin[1,2,3], Elisabete M. Nascimento[1,2,3], Chandresh R. Gajera[1,2,3], Lu Chen[1,2,3], Patrick Neuhöfer[1,2,3], Alina Garbuzov[1,2,3], Sui Wang[4] & Steven E. Artandi[1,2,3]*

Hepatocytes are replenished gradually during homeostasis and robustly after liver injury[1,2]. In adults, new hepatocytes originate from the existing hepatocyte pool[3–8], but the cellular source of renewing hepatocytes remains unclear. Telomerase is expressed in many stem cell populations, and mutations in telomerase pathway genes have been linked to liver diseases[9–11]. Here we identify a subset of hepatocytes that expresses high levels of telomerase and show that this hepatocyte subset repopulates the liver during homeostasis and injury. Using lineage tracing from the telomerase reverse transcriptase (*Tert*) locus in mice, we demonstrate that rare hepatocytes with high telomerase expression (TERT[High] hepatocytes) are distributed throughout the liver lobule. During homeostasis, these cells regenerate hepatocytes in all lobular zones, and both self-renew and differentiate to yield expanding hepatocyte clones that eventually dominate the liver. In response to injury, the repopulating activity of TERT[High] hepatocytes is accelerated and their progeny cross zonal boundaries. RNA sequencing shows that metabolic genes are downregulated in TERT[High] hepatocytes, indicating that metabolic activity and repopulating activity may be segregated within the hepatocyte lineage. Genetic ablation of TERT[High] hepatocytes combined with chemical injury causes a marked increase in stellate cell activation and fibrosis. These results provide support for a 'distributed model' of hepatocyte renewal in which a subset of hepatocytes dispersed throughout the lobule clonally expands to maintain liver mass.

Hepatocytes execute the metabolic activities of the liver and show functional heterogeneity along the axis within the lobule defined from the portal vein to the central vein[12]. At the extreme ends of this axis, pericentral AXIN2[+] hepatocytes repopulate the liver during normal homeostasis[13], whereas periportal hepatocytes marked by SOX9 expression are inactive during homeostasis but expand in response to chronic chemical damage[14]. Observations indicating that proliferating hepatocytes are located throughout the lobule[15,16] suggest that additional sources of repopulating hepatocytes exist. Telomerase synthesizes telomere repeats and has been linked to long-term renewal in stem cells and cancers[17]. Germline inactivating mutations in telomerase genes predispose humans and mice to cirrhosis[9–11], while activating mutations in the *TERT* promoter represent the most recurrent mutations in hepatocellular carcinoma[18]. Given the important roles of telomerase in liver disease, and observations that telomerase is found in stem cell compartments in multiple adult tissues[19–22], we hypothesized that telomerase may be expressed in liver cells with unique properties.

To identify telomerase-expressing cells in vivo, we engineered a mouse strain expressing the inducible CreER[T2] recombinase from the endogenous *Tert* locus (Extended Data Fig. 1a–d). Treatment of *Tert*[CreERT2/+] knock-in mouse embryonic stem (ES) cells in culture with 4-hydroxy tamoxifen resulted in efficient recombination of a fluorescent reporter (Extended Data Fig. 1e–g). To study the adult liver, we crossed *Tert*[CreERT2/+] mice with a *Rosa26*[LSL-Tomato/+] reporter strain that enables permanent cell labelling by deletion of a transcriptional

stop element flanked by *loxP* sites and concomitant expression of fluorescent Tomato protein. *Tert*[CreERT2/+]*Rosa26*[LSL-Tomato/+] mice were injected with a near-saturating dose of tamoxifen (1 mg per 10 g body weight; Extended Data Fig. 1i), and analysed 3 days later (Fig. 1a). We found that a subset of cells throughout the liver expressed Tomato and the hepatocyte marker HNF4A (Fig. 1b). Tomato expression was not detected in other liver cell types (Extended Data Fig. 1k–n). To isolate these TERT[High] hepatocytes by fluorescence-activated cell sorting (FACS) (see Supplementary Information), we labelled all hepatocytes with an adeno-associated virus expressing hepatocyte-specific GFP (AAV-GFP)[23] (Fig. 1c and Extended Data Fig. 1h). We found that all
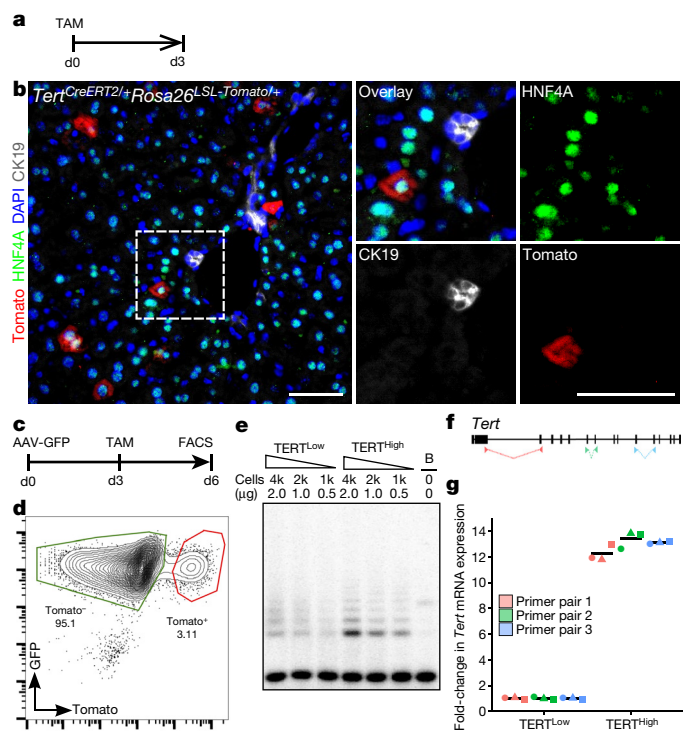
Fig. 1 | Identification of a hepatocyte subpopulation with elevated *Tert* and telomerase activity. a, **b**, Immunofluorescence analysis of *Tert*[CreERT2/+]*Rosa26*[LSL-Tomato/+] livers treated with a single dose of tamoxifen and analysed 3 days later (timeline shown in **a**; d, day). Tomato (red), HNF4A (green), CK19 (white), and DAPI (blue) shown. **c–g**, Analysis of telomerase expression in FACS-sorted hepatocytes. **c**, Timeline. **d**, Representative FACS plot. **e**, TRAP assay (B, buffer only). **f**, Position of primer pairs for RT–qPCR. **g**, Fold-change in *Tert* mRNA expression between TERT[High] and TERT[Low] hepatocytes. *n* = 3 mice, each indicated by a unique shape; horizontal bars show mean. Experiments repeated three times for **b**, more than five times for **d**, and twice for **e** and **g**. Scale bars, 50 μm.

[1]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [3]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Ophthalmology, Stanford University School of Medicine, Stanford, CA, USA. *e-mail: sartandi@stanford.edu
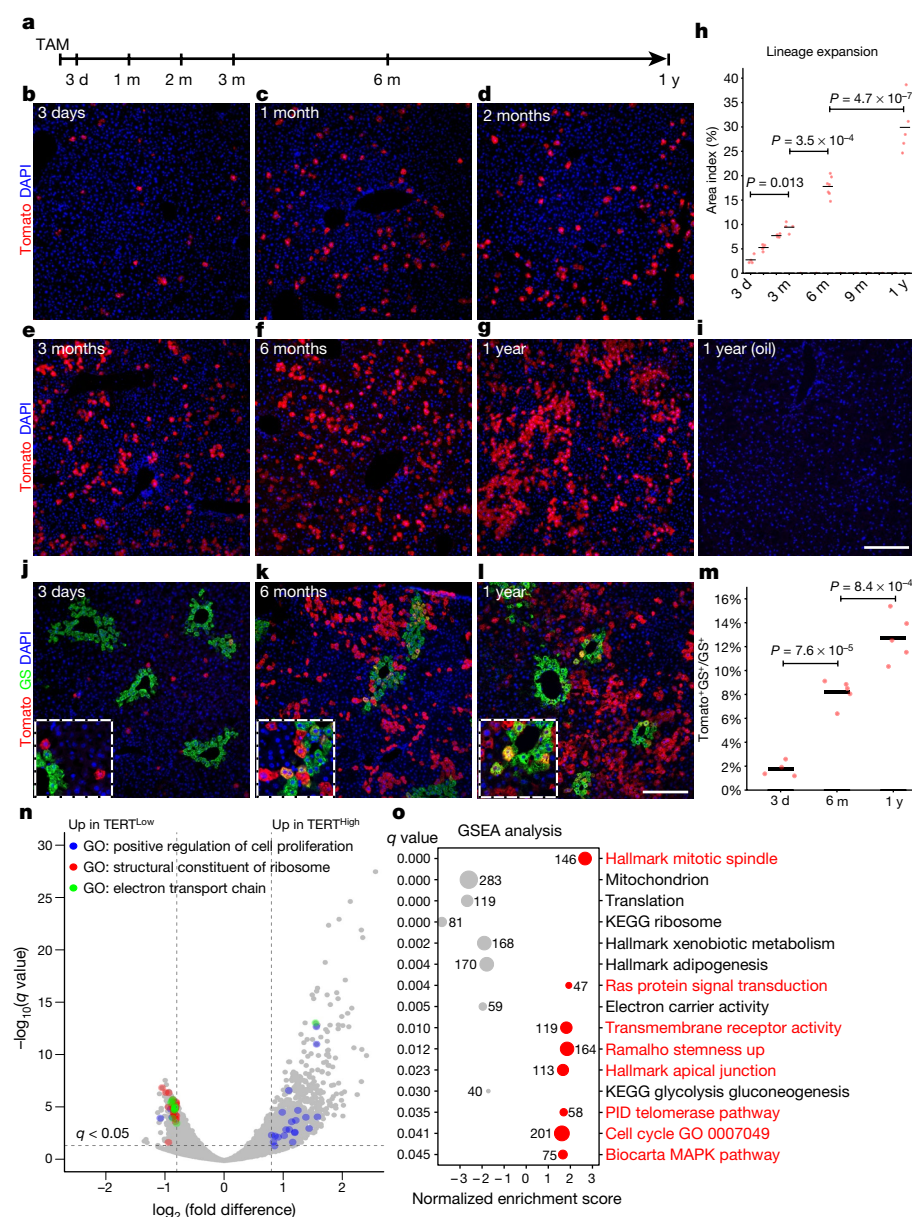
Fig. 2 | TERT^{High} hepatocytes repopulate the liver in homeostasis and show downregulation of metabolic genes. a–i, Lineage tracing in $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice treated with single-dose tamoxifen (b–g) or oil vehicle (i) and analysed after indicated tracing periods by immunofluorescence for Tomato. Timeline shown in a; d, days; m, months; y, years. h, Quantification of Tomato$^+$ hepatocyte area ($n = 4, 5, 4, 7, 5$ mice for each time-point from left to right; horizontal bars show mean). j–l, Co-immunofluorescence for Tomato and GS in $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice traced for 3 days, 6 months or 1 year. m, Quantification of Tomato$^+$GS$^+$ fraction of GS$^+$ hepatocytes ($n = 4, 5, 5$ mice for each time-point; horizontal bars show mean). n, o, RNA-seq results for FACS-purified TERT^{High} (Tomato$^+$) or TERT^{Low} (Tomato$^-$) hepatocytes ($n = 3$ mice for each group). n, Volcano plot for enriched genes and Gene Ontology (GO) terms (cut-offs: $q < 0.05$, $|\log_2(\text{fold difference})| > 0.8$). o, Gene set enrichment analysis (GSEA) for enriched gene-sets (number of genes shown for each gene set). Red, enriched in TERT^{High} cells; grey, enriched in TERT^{Low} cells. Experiments repeated twice for time-points in b, f, g, j–l. Scale bars, 100 μm.

Tomato$^+$ cells were also GFP$^+$, typically representing 3–5% of all hepatocytes from 2-month-old mice (Fig. 1d). Telomeric repeat amplification protocol (TRAP) showed a fivefold increase in telomerase activity in the TERT^{High} population (GFP$^+$Tomato$^+$) compared with the TERT^{Low} population (GFP$^+$Tomato$^-$) (Fig. 1e, Extended Data Fig. 1j and Supplementary Fig. 1). Quantitative reverse transcription PCR (RT–qPCR) showed that there was 12.9-fold more $Tert$ mRNA in the TERT^{High} population than in the bulk TERT^{Low} hepatocyte population (Fig. 1f, g). Both populations comprised a similar distribution of diploid and polyploid cells (Extended Data Fig. 2). These data show that $Tert$ mRNA and telomerase activity are elevated in TERT^{High} hepatocytes.

To determine whether TERT^{High} hepatocytes repopulate the liver during homeostasis, we performed lineage tracing by injecting two-month old $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice with a single dose of tamoxifen (1 mg per 10 g body weight) or oil vehicle and allowed these animals to age for up to 1 year (Fig. 2a). TERT^{High} hepatocytes represented $2.8 \pm 0.4\%$ of liver area 3 days after tamoxifen treatment, but the Tomato$^+$ progeny of these cells increased progressively during the tracing period to comprise $29.9 \pm 2.4\%$ of liver area at 1 year (Fig. 2b–h). All Tomato$^+$ cells detected after 1 year were HNF4A$^+$ hepatocytes (Extended Data Fig. 3a–e), and Tomato$^+$ cells were not detected in mice treated with oil vehicle (Fig. 2i). A single tamoxifen

injection generated a similar number of Tomato$^+$ hepatocytes as three injections administered at 5-week intervals over the same tracing period (Extended Data Fig. 3f–i), indicating that elevated $Tert$ promoter activity is an intrinsic feature of cell identity. Co-staining of sections from this lineage tracing time course series for Tomato and the pericentral zone marker glutamine synthetase (GS)[12] showed that TERT^{High} hepatocytes were distributed throughout all lobular zones. The vast majority of TERT^{High} hepatocytes were located in the periportal and midlobular zones (3-day trace), and the progeny from these cells expanded markedly to replenish hepatocytes in these zones. Within the pericentral zone, the TERT^{High} lineage comprised $1.8 \pm 0.3\%$ of cells at 3 days (Fig. 2j), but increased over time ($8.2 \pm 0.5\%$ at 6 months and $12.7 \pm 0.9\%$ at 1 year; Fig. 2k–m and Extended Data Fig. 4). Analysis of proliferating hepatocyte position by Ki-67 immunostaining revealed that Ki-67$^+$ hepatocytes were dispersed throughout all lobular zones in both wild-type and $Tert^{CreERT2/+}$ mice, matching the distributed pattern of TERT^{High} hepatocytes (Extended Data Fig. 5a, b, h). These data show that rare TERT^{High} hepatocytes drive a marked and progressive repopulation of the hepatocyte lineage throughout the lobule during normal homeostasis.

To understand how TERT^{High} cells differ from bulk hepatocytes, we performed RNA sequencing (RNA-seq) on TERT^{High} and TERT^{Low}
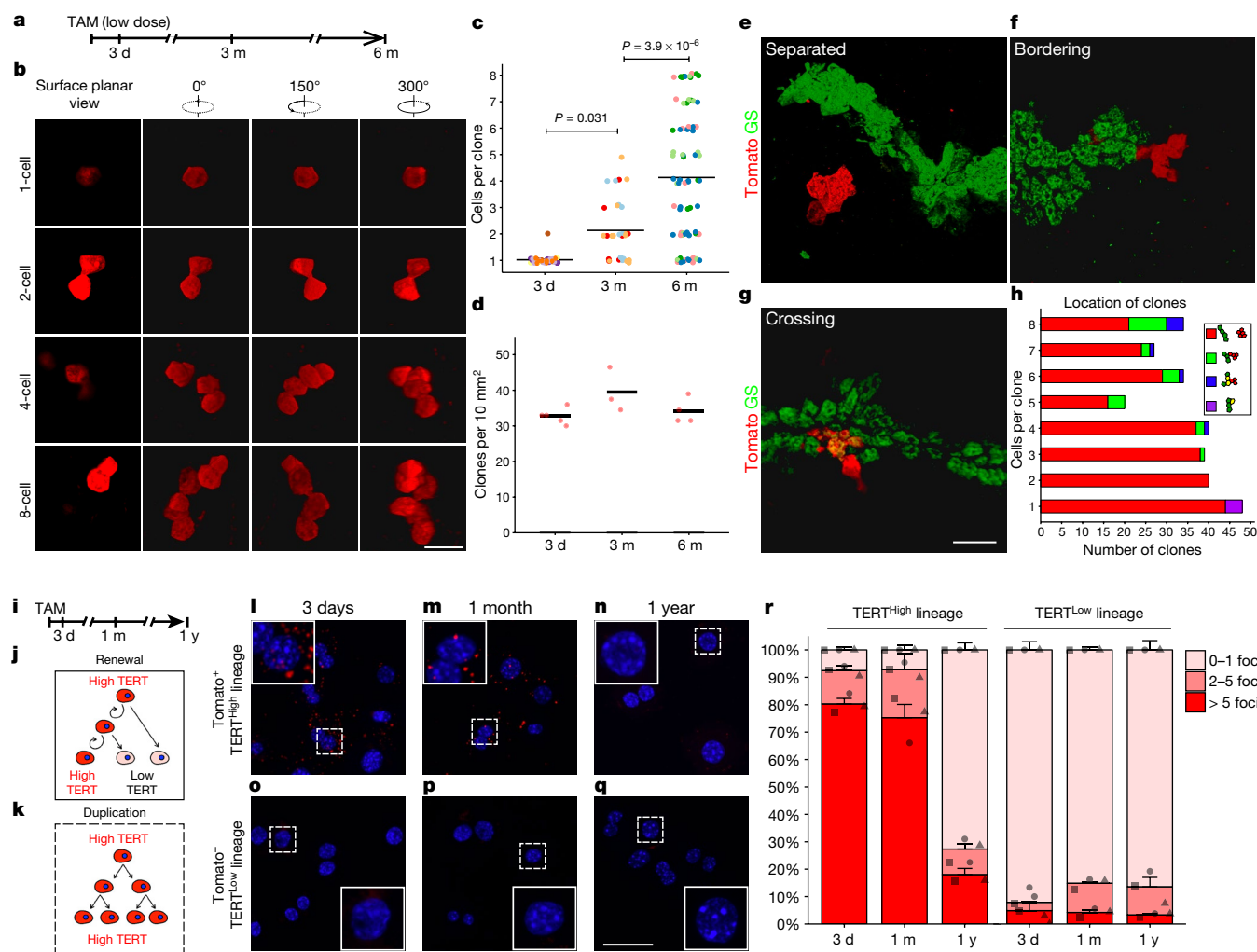
Fig. 3 | TERT^High hepatocytes drive clonal expansion by a self-renewal mechanism. a–h, 3D analysis of sparse labelled hepatocyte clones in $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice treated with low-dose tamoxifen and traced for 3 days, 3 months or 6 months. **a,** Timeline. **b–d,** Clone sizes (**b, c**) and clone number per volume (**d**) ($n = 5$, 3, and 4 mice, respectively; each mouse represented by unique dot colour in **c**; horizontal bars show mean). **e–g,** Co-immunofluorescence for Tomato and GS to assess clone location in 6-month trace samples. **h,** Quantification of clone size and position relative to GS^+ cells. **i–r,** Single-molecule RNA FISH on FACS-purified TERT^High derived (**l–n**) and TERT^Low derived (**o–q**) hepatocytes. **r,** Quantification by number of foci ($n = 3$ mice per time point, each mouse indicated by unique shape). Bars and error bars are mean + s.e.m. For cells with more than five foci, $P_{1month-3days} = 0.56$, $P_{1year-3days} = 2.7 \times 10^{-5}$, $P_{1year-1month} = 4.3 \times 10^{-5}$). Experiments repeated twice. Scale bars, 50 μm.

hepatocytes isolated by FACS from three $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice, 3 days after tamoxifen treatment. RNA-seq identified 3,172 genes that were differentially expressed between the two populations ($q < 0.05$; Fig. 2n, Extended Data Fig. 3j). Gene Ontology analysis (Fig. 2n) and the Database for Annotation, Visualization and Integrated Discovery (DAVID, Extended Data Fig. 3j) showed that cell cycle genes were upregulated in the TERT^High population, whereas ribosomal genes and mitochondrial genes were upregulated in the TERT^Low population. Gene set enrichment analysis (GSEA) revealed increased representation of gene sets associated with cell division and receptor tyrosine kinase activity in the TERT^High population (Fig. 2o, red), and decreased representation of gene sets associated with ribosome components, mitochondrial proteins, the electron transport chain and hepatocyte metabolic activities (Fig. 2o, grey). Proliferation in TERT^High hepatocytes was higher than in TERT^Low hepatocytes ($6.4 \pm 1.0\%$ versus $0.9 \pm 0.1\%$) by 5-ethynyl-2′-deoxyuridine (EdU) incorporation (for 7 days in drinking water, Extended Data Fig. 6). Together, these data suggest that TERT^High hepatocytes are less invested in the metabolic and synthetic functions of bulk hepatocytes, and more dedicated to proliferation and homeostatic renewal, than TERT^Low hepatocytes.

To characterize the behaviour of single TERT^High hepatocytes and their progeny through clonal analysis and sparse labelling, we injected

$Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ mice with a lower dose of tamoxifen (0.08 mg per 10 g body weight), and performed lineage tracing for 3 days, 3 months or 6 months (Fig. 3a). Confocal microscopy was performed on thick tissue sections and followed by three-dimensional reconstruction (Fig. 3b). The average clone size increased progressively from single cells at 3 days to 2.1 ± 0.2 cells at 3 months and 4.2 ± 0.4 cells at 6 months (Fig. 3c). Average clonal density did not change, indicating that there was no significant loss of TERT^High hepatocyte clones over the 6-month trace (Fig. 3d). The irregular shape of these clones matches the anatomical organization of hepatocytes within hepatic cords[24]. Co-staining of 6-month trace samples with antibodies against GS revealed that the vast majority of clones resided outside the GS^+ zone (Fig. 3e, h (red bars)), and a subset of these bordered the GS^+ peri-central zone (Fig. 3f, h (green bars)). We also found occasional clones comprised of a mixture of GS^+ and GS^− cells (Fig. 3g, h (blue bars)). The 'cross-zone' clones derive from TERT^High hepatocytes but comprise cells with two distinct zonal fates. These clonal studies matched the findings on homeostatic expansion of the TERT^High lineage (Fig. 2), and further supported the idea that TERT^High hepatocytes are a key source of hepatocyte renewal.

TERT^High hepatocytes could generate clones either by a self-renewal and differentiation mechanism, in which daughter cells are comprised
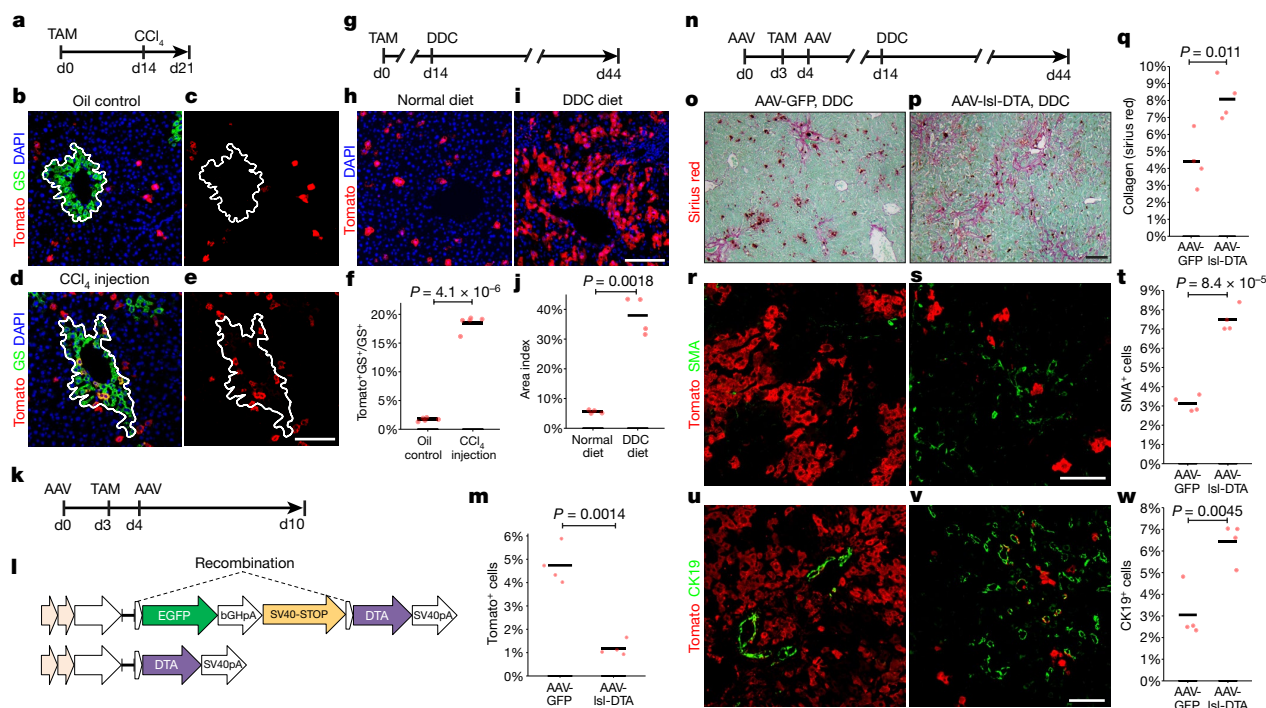
Fig. 4 | TERT^High hepatocytes are critical for liver regeneration.
a–f, Single-dose CCl₄-induced liver injury. **a**, Timeline.
b–e, Immunofluorescence for Tomato and GS of cells from
Tert^CreERT2/+Rosa26^LSL-Tomato/+ mice treated with oil vehicle (**b**, **c**) or
CCl₄ (**d**, **e**) at day 7 post-treatment. White lines show GS⁺ pericentral
zone. **f**, Quantification (*n* = 5 mice). **g–j**, DDC diet-induced injury in
Tert^CreERT2/+Rosa26^LSL-Tomato/+ mice. **g**, Timeline. **h, i**, Immunofluorescence
of liver for Tomato after 30 days treatment with normal diet (**h**) or
DDC diet (**i**). **j**, Quantification (*n* = 4 mice). **k–m**, Ablation of TERT^High

hepatocytes via AAV-lsl-DTA injection into *Tert^CreERT2/+Rosa26^LSL-Tomato/+*
mice (*n* = 4 mice). **k**, Timeline. **l**, The AAV-lsl-DTA construct.
m, Quantification of Tomato⁺ cells. **n–v**, Genetic ablation followed
by DDC injury in *Tert^CreERT2/+Rosa26^LSL-Tomato/+* mice. **n**, Timeline.
o–w, Livers analysed by SiriusRed for collagen (**o–q**), SMA for activated
stellate cells (**r–t**), and CK19 (**u–w**) (*n* = 4 mice). **o, r, t**, AAV-GFP injected
control animals; **p, s, u**, AAV-lsl-DTA-injected animals. Horizontal bars
show mean. Experiments repeated at least twice. Scale bars, 100 μm.

of both TERT^High and TERT^Low cells (Fig. 3j), or by a simple duplica-
tion mechanism, in which all daughter cells remain TERT^High (Fig. 3k).
To distinguish between these possibilities, we examined *Tert* mRNA
with single-molecule RNA fluorescence in situ hybridization (FISH)
on sorted Tomato⁺ and Tomato⁻ hepatocytes from different trac-
ing periods (Fig. 3i, l–r), as well as wild-type hepatocytes (Extended
Data Fig. 7). The percentage of Tomato⁺ cells with high *Tert* mRNA
(more than five mRNA foci) was comparable at 3 days and 1 month
(80.3 ± 2.0% versus 75.3 ± 4.8%), but decreased to 18.0 ± 2.2% after
1 year. Tomato⁻ cells remained TERT^Low, regardless of the tracing period.
The presence of rare cells in this fraction with high *Tert* mRNA
is likely to indicate incomplete recombination with CreER^T2. These studies indi-
cate that the TERT^High subpopulation both self-renews to replenish
TERT^High cells and differentiates to yield TERT^Low daughter cells.

To understand the ability of TERT^High hepatocytes to replace dam-
aged cells in the pericentral zone, we eliminated pericentral hepatocytes
by single-dose carbon tetrachloride (CCl₄) injection[25] (Extended Data
Fig. 8c–f). Although TERT^High hepatocytes are rare within the GS⁺ peri-
central zone, there was a marked increase in the number of GS⁺Tomato⁺
cells 7 days after injury (Fig. 4a–f). These data indicate that injury to
pericentral hepatocytes activates nearby TERT^High hepatocytes, and
that their progeny assume a new zonal identity in healing pericentral
wounds. To understand whether TERT^High hepatocytes contribute to
hepatocyte regeneration after global injury, we challenged the livers with
a 0.1% 3,5-diethoxycarbonyl-1,4-dihydrocollidine (DDC) diet (Fig. 4g
and Extended Data Fig. 8g, h). There was a significant expansion of
Tomato⁺ hepatocytes after 1 month of the DDC diet (38.0 ± 3.2% ver-
sus 5.6 ± 0.3% in control livers, *P* = 0.0018; Fig. 4g–j). Some progeny of
TERT^High cells adopted a ductal fate (Extended Data Fig. 9), consistent
with known hepatocyte plasticity following DDC-induced injury[26].
These findings reveal that TERT^High hepatocytes repopulate hepatocytes
at an accelerated rate in the setting of chemical injury.

To determine whether TERT^High hepatocytes are required for nor-
mal injury responses, we ablated *Tert*-expressing hepatocytes using
a diphtheria toxin (DTA)-based AAV system, in which hepatocyte-
specific expression of DTA is induced upon Cre-mediated deletion
of a *loxP-EGFP-Stop-loxP* element (Fig. 4k–m). Intravenous infection
of wild-type mice with both AAV-lsl-DTA and AAV-Cre resulted in
massive hepatocyte necrosis and death within 6 days, whereas infection
with AAV-lsl-DTA alone was well tolerated for up to 2 months and did
not induce liver damage (Extended Data Fig. 10f–j). When we used
this system in *Tert^CreERT2/+Rosa26^LSL-Tomato/+* mice, the abundance of
TERT^High (Tomato⁺) cells was reduced by 75.1% in mice treated with
AAV-lsl-DTA compared with those treated with AAV-GFP (Fig. 4m).
After ablating TERT^High cells, we induced liver injury with a DDC diet
for 30 days (Fig. 4n). Expansion of the TERT^High (Tomato⁺) cell lin-
eage was substantially suppressed in mice treated with AAV-lsl-DTA
(Fig. 4s, v) compared with those treated with AAV-GFP (Fig. 4r, u).
DDC treatment following ablation of TERT^High hepatocytes led to a
marked increase in liver fibrosis, shown by an increase in collagen
deposition (Fig. 4o–q) and a concomitant increase in the number of
activated stellate cells (Fig. 4r–t). There was an associated increase in
cells positive for the ductal marker CK19 (Fig. 4u–w), indicating that
suppression of hepatocyte renewal enhances the ductal reaction char-
acteristic of DDC treatment. Finally, we replicated these results using an
independently constructed AAV that allows induction of DTA through
Cre-mediated inversion and deletion steps (AAV-flex-DTA) (Extended
Data Fig. 10). Together, these data show that TERT^High hepatocytes
are critical for normal liver regeneration in the setting of DDC injury
and that regeneration in their absence results in elevated stellate cell
activation and fibrosis.

On the basis of the dispersed location of TERT^High hepatocytes and
their clonal behaviour during regeneration, we propose a 'distributed
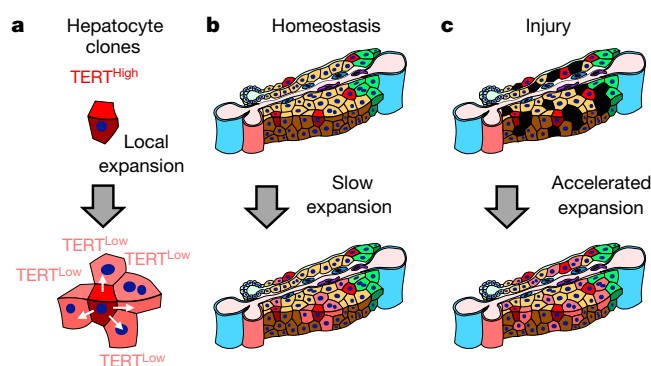model' to explain hepatocyte renewal. According to this model, rare

Fig. 5 | A distributed model to explain hepatocyte renewal. a, TERT[High] hepatocytes generate expanding clones comprising both TERT[High] and TERT[Low] hepatocytes. **b**, Rare TERT[High] hepatocytes distributed throughout the lobule form enlarging clones during homeostasis in response to local hepatocyte loss. **c**, Repopulation of injured liver by clonal expansion of TERT[High] hepatocytes is accelerated during injury. Blue vessels, portal and central veins; pink vessel, hepatic artery; small green cells, cholangiocytes; large green cells, pericentral hepatocytes; red cells, TERT[High] hepatocytes; black cells, damaged hepatocytes.

TERT[High] hepatocytes located throughout the lobule form enlarging clones during homeostasis in response to hepatocyte loss, and this response is accelerated during liver injury (Fig. 5). These findings provide a framework that can explain several longstanding observations in hepatocyte renewal, including the ability of the liver to recover from injuries in any lobular zone; a general lack of evidence for long-range migration of hepatocytes; and the presence of rare proliferating hepatocytes throughout the lobule. Our RNA-seq data suggest that repopulating activity and metabolism may be segregated within the hepatocyte population. Telomerase activity is critical for preserving long-term cell division and chromosomal stability. Maintaining the liver using a subset of hepatocytes with elevated telomerase and reduced metabolic activity may be important for long-term tissue maintenance, for preventing the accrual of damaged DNA caused by reactive oxygen species and for suppressing hepatocellular carcinoma. We speculate that depletion or dysfunction of an analogous subset of repopulating hepatocytes in humans may underlie the pathophysiology of cirrhosis. Strategies to mitigate this cellular depletion may prove useful in treating cirrhosis of diverse aetiologies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0004-7.

1. Stanger, B. Z. Cellular homeostasis and repair in the mammalian liver. *Annu. Rev. Physiol* **77**, 179–200 (2015).
2. Zaret, K. S. & Grompe, M. Generation and regeneration of cells of the liver and pancreas. *Science* **322**, 1490–1494 (2008).
3. Malato, Y. et al. Fate tracing of mature hepatocytes in mouse liver homeostasis and regeneration. *J. Clin. Invest.* **121**, 4850–4860 (2011).
4. Schaub, J. R., Malato, Y., Gormond, C. & Willenbring, H. Evidence against a stem cell origin of new hepatocytes in a common mouse model of chronic liver injury. *Cell Reports* **8**, 933–939 (2014).
5. Yanger, K. et al. Adult hepatocytes are generated by self-duplication rather than stem cell differentiation. *Cell Stem Cell* **15**, 340–349 (2014).
6. Tarlow, B. D., Finegold, M. J. & Grompe, M. Clonal tracing of Sox9+ liver progenitors in mouse oval cell injury. *Hepatology* **60**, 278–289 (2014).
7. Tarlow, B. D. et al. Bipotential adult liver progenitors are derived from chronically injured mature hepatocytes. *Cell Stem Cell* **15**, 605–618 (2014).
8. Raven, A. et al. Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. *Nature* **547**, 350–354 (2017).
9. Calado, R. T. et al. Constitutional telomerase mutations are genetic risk factors for cirrhosis. *Hepatology* **53**, 1600–1607 (2011).
10. Hartmann, D. et al. Telomerase gene mutations are associated with cirrhosis formation. *Hepatology* **53**, 1608–1617 (2011).
11. Rudolph, K. L., Chang, S., Millard, M., Schreiber-Agus, N. & DePinho, R. A. Inhibition of experimental liver cirrhosis in mice by telomerase gene delivery. *Science* **287**, 1253–1258 (2000).
12. Jungermann, K. & Kietzmann, T. Zonation of parenchymal and nonparenchymal metabolism in liver. *Annu. Rev. Nutr* **16**, 179–203 (1996).
13. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
14. Font-Burgada, J. et al. Hybrid periportal hepatocytes regenerate the injured liver without giving rise to cancer. *Cell* **162**, 766–779 (2015).
15. Planas-Paz, L. et al. The RSPO-LGR4/5-ZNRF3/RNF43 module controls liver zonation and size. *Nat. Cell Biol.* **18**, 467–479 (2016).
16. Tanami, S. et al. Dynamic zonation of liver polyploidy. *Cell Tissue Res* **368**, 405–410 (2017).
17. Batista, L. F. & Artandi, S. E. Understanding telomere diseases through analysis of patient-derived iPS cells. *Curr. Opin. Genet. Dev* **23**, 526–533 (2013).
18. Nault, J. C. et al. High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat. Commun.* **4**, 2218 (2013).
19. Pech, M. F. et al. High telomerase is a hallmark of undifferentiated spermatogonia and is required for maintenance of male germline stem cells. *Genes Dev.* **29**, 2420–2434 (2015).
20. Montgomery, R. K. et al. Mouse telomerase reverse transcriptase (mTert) expression marks slowly cycling intestinal stem cells. *Proc. Natl Acad. Sci. USA* **108**, 179–184 (2011).
21. Schepers, A. G., Vries, R., van den Born, M., van de Wetering, M. & Clevers, H. Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *EMBO J* **30**, 1104–1109 (2011).
22. Itzkovitz, S. et al. Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nat. Cell Biol.* **14**, 106–114 (2011).
23. Gao, G.-P. et al. Novel adeno-associated viruses from rhesus monkeys as vectors for human gene therapy. *Proc. Natl Acad. Sci. USA* **99**, 11854–11859 (2002).
24. Elias, H. A re-examination of the structure of the mammalian liver; the hepatic lobule and its relation to the vascular and biliary systems. *Am. J. Anat.* **85**, 379–456 (1949).
25. Recknagel, R. O., Glende, E. A. Jr, Dolak, J. A. & Waller, R. L. Mechanisms of carbon tetrachloride toxicity. *Pharmacol. Ther.* **43**, 139–154 (1989).
26. Kopp, J. L., Grompe, M. & Sander, M. Stem cells versus plasticity in liver and pancreas regeneration. *Nat. Cell Biol.* **18**, 238–245 (2016).

Author contributions S.L. and S.E.A. conceived the study. S.L., E.M.N. and S.E.A. designed the experiments. S.L. and C.R.G. created the *Tert* knock-in line. S.L. and E.M.N. performed the lineage tracing and EdU incorporation experiments. L.C. performed the TRAP assay. S.L. and P.N. performed histological analysis. S.L. and S.W. performed the AAV experiments. S.L. and A.G. performed RNA-seq analyses. S.L. and S.E.A. analysed the data and wrote the paper.

METHODS

Generation of the *Tert*[CreERT2] knock-in line. The targeting vector was generated by serial recombineering and gate-way cloning. Homology arms (mm10 chr13: 73,621,344–73,631,102) were cloned from the BAC (RP24-342O18) via recombineering. A codon-optimized intron-CreERT2-NeoR cassette[27] was inserted into the endogenous translational start site of *Tert* (mm10 chr13: 73,627,032–73,627,033) via recombineering. The final targeting vector was created via gate-way cloning to the pWS-TK2 vector with thymidine kinase cassettes at both ends of the homology arms, as previously described[28]. The targeting vector was linearized and electroporated into JM8/F6 mouse ES cells. Correctly targeted ES clones were selected by Southern blotting and karyotypes, and then injected into ICR/CD-1 blastocysts to generate the knock-in line. *Tert*[CreERT2/+] mice were born at normal Mendelian frequency. To verify the efficacy of CreERT2 in the ES cells, the *Tert*[CreERT2/+] clone was targeted with a modified Rosa26-mTmG targeting vector[29] using HygroR as the selection gene. The double knock-in cells were treated with 500 nM 4-hydroxy tamoxifen (4-OHT) to evaluate recombination efficiency.

AAV production. All AAVs used in this study were produced with *cis*-plasmids containing the full TBG promoter (two copies of the α-1-microglobulin/bikunin precursor (*AMBP*) enhancer elements followed by the promoter of the *SERPINA7* gene and a mini-intron), an AAV8 serotype packaging plasmid, and an adenovirus helper plasmid. AAV-GFP (AAV8-TBG-PI-eGFP-WPRE-bGH, catalogue no. AV-8-PV0146) and AAV-Cre (AAV8-TBG-PI-Cre-rBG, catalogue no. AV-8-PV1091) were purchased from the University of Pennsylvania Vector Core. AAV-lsl-DTA contains a strong SV40 stop element cloned from the Lox-Stop-Lox TOPO plasmid[30] (addgene Plasmid no. 11584). AAV-flex-DTA was modified from pAAV-mCherry-flex-DTA[31] (addgene Plasmid no. 58536) with the following changes: the EF-1α promoter was swapped with the TBG promoter, and mCherry was swapped with EGFP. HEK293T cells were transfected and grown on Corning multi-layer flasks to produce the viruses. The viruses were purified by Iodixanol (Sigma-Aldrich) gradient ultracentrifugation[32], and titred by qPCR[33] and SYPRO Ruby (ThermoFisher) protein gel staining with standards.

Animals. *Tert*[CreERT2/+] mice were bred with the Rosa26 reporter (*Gt(ROSA)26Sortm14(CAG-tdTomato)Hze/J*) line[34] to generate *Tert*[CreERT2/+]*Rosa26*[LSL-Tomato/+] mice for analysis. Two-month-old mice were intraperitoneally injected with tamoxifen (Caymon, 1 mg per 10 g weight) dissolved in 100 μl sesame oil (Sigma-Aldrich). Sparse labelling was achieved by injecting tamoxifen at 0.08 mg per 10 g weight. EdU (Carbosynth) was administrated via drinking water (1 mg/ml) daily for seven days. AAV was diluted to 4×10^{11} genome particles in 100 μl normal saline (per mouse), and injected intravenously. For DDC injury, mice received diet TD.07571 (Harlan) containing 0.1% DDC (Sigma-Aldrich) ad libitum. For CCl$_4$ injury, mice were injected with liquid CCl$_4$ (Sigma-Aldrich, 10 μl per 10 g weight) dissolved in sesame oil (Sigma-Aldrich).

Statistics. No statistical methods were used to predetermine sample sizes. When comparing two groups, *P* values were determined by two-sided unpaired *t*-test. When comparing more than two groups, *P* values were determined by one-way ANOVA with Tukey's HSD test performed as the post hoc analysis. Data significance was also tested by non-parametric statistics using two-sided unpaired Wilcoxon–Mann–Whitney test for two-group comparison, and Kruskal–Wallis one-way ANOVA on ranks with Conover–Iman test performed as the post hoc analysis for more than two groups. Kolmogorov–Smirnov test was performed to compare the distribution patterns of continuous variables. The animals were randomly assigned to each experimental or control group. The investigators were not blinded to allocation during experiments and outcome assessment. Data are presented as mean ± s.e.m. Graphs were generated by the ggplot2 package[35] in R.

FACS experiments. Cells were isolated by standard two-step collagenase perfusion. Liver perfusion medium (Life Technologies) and filtered (0.22 μm) liver digest medium (Life Technologies) were perfused via the portal vein sequentially, according to the manufacturer's instructions. Dissociated liver was passed through a 100-μm cell strainer and the hepatocytes were enriched by low-speed centrifugation (50*g* for 3 min) three times in hepatocyte wash medium (Life Technologies). Cells were analysed and/or sorted with a BD Aria II flow cytometer using a 100-μm nozzle. Dead cells were excluded by Topro3 (1 μM) or DAPI (1 μM) (Life Technologies) incorporation. For ploidy analysis, hepatocytes were incubated in Hoechst 33342 (15 μg/ml) and Reserpine (5 μM) at 37 °C for 30 min before analysis.

Immunofluorescence, imunohistochemistry, EdU detection, single-molecule RNA FISH and SiriusRed staining. Livers were cut into small blocks, and fixed in zinc-buffered formalin (Anatech). For immunofluorescence, tissue blocks were fixed overnight at 4 °C, cryoprotected in 30% (w/v) sucrose, embedded in OCT, snap-frozen and cut into 7-μm cryosections. For thick tissue analysis, tissue blocks were briefly fixed, embedded in low-melting agarose and cut into 300-μm sections using a vibratome, as previously described[36]. For immunohistochemistry, tissue blocks were fixed overnight at 4 °C, incubated in 70% ethanol overnight, embedded in paraffin and cut into 5-μm sections. Antigen retrieval was performed with citrate (pH 6) buffer (Biogenex) for 10 min using a pressure cooker. Slides were stained with primary and secondary antibodies in blocking buffer (1% BSA, 5% donkey serum, 0.25% Triton-X in PBS) overnight at 4 °C, incubated with 1 mM DAPI for 5 min at room temperature, and mounted in Aqua poly/mount (Polysciences), or Vectashield with DAPI (Vector laboratories). DAB Peroxidase Substrate Kit (Vector Laboratories) or Emerald chromogen kit (Abcam) were used for immunohistochemistry. EdU incorporation was detected by using the Click-iT EdU Alexa Fluor 488 Imaging Kit (Life Technologies). For analysis on cytospun samples, cells were FACS-sorted and cytospun (500 r.p.m./28*g* for 5 min) onto slides. Slides were fixed in 4% (v/v) PFA for 5 min, and stained with primary and secondary antibodies in blocking buffer for 1 h at room temperature, and then mounted in Prolong Gold with DAPI mounting medium (Life Technologies). Alternatively, slides were fixed in 4% (v/v) PFA for 20 min, and processed for single-molecule RNA FISH using an RNAscope 2.0 HD Detection-RED kit (ACDbio) according to the manufacturer's instructions. SiriusRed staining for collagen deposit was performed with Fast Green as the counter-stain, using a staining kit (Chondrex), according to the manufacturer's instructions.

RT–qPCR and RNA-seq. RT–qPCR and RNA-seq were performed on TERT[High] and TERT[Low] hepatocytes isolated by FACS from three *Tert*[CreERT2/+]*Rosa26*[LSL-Tomato/+] mice three days after tamoxifen treatment. Hepatocytes were sorted directly in TRIzol-LS (Life Technologies). Total RNA was extracted and purified using an RNeasy micro kit (Qiagen) according to the manufacturer's instructions. qPCR was performed using the following primers: *Tert* (pair1)[37] CCACGTATGTGTCCATCAGC/TAGAGGATTGCCACTGGCTC; *Tert* (pair2) ATCTGCAGGATTCAGATGCC/GCAGGAAGTGCAGGAAGAAG; *Tert* (pair3)[21] TGGCTTGCTGCTGGACACTC/TGAGGCTCGTCTTAATTGAGGTCTG; *Gtf2b* CTCTGTGGCGGCAGCAGCTATTT/CGAGGGTAGATCAGTCTGTAGGA. qPCR reactions were carried out using Brilliant II SYBR Green master mix (Strategene) and Roche lightcycler 480. Quantitation cycle (Cq) values were determined by the second derivative maximum method, and fold-changes were calculated by $2^{-\Delta Cq}$. RNA-seq libraries were constructed using a KAPA Stranded mRNA-Seq Kit (Kapa). Libraries were sequenced on the Illumina NextSeq platform, generating about 55–75 million 75-bp paired-end reads per library. Three biological replicates per sample were analysed. Raw reads were trimmed by TrimGalore 0.4.0 (Babraham Bioinformatics), mapped to mm10 by tophat 2.0.13[38], analysed by the DEseq2 packages[39].

Imaging analysis. Fluorescent images were analysed using Leica LAS AF, ImageJ, Adobe Photoshop and Fluorender. Area index was defined by the liver area covered by Tomato$^+$ cells as the percentage of total area, and quantified by ImageJ. 3D reconstruction was performed using Fluorender. Multicellular clones were imaged in $258 \times 258 \times 100$-μm^3 volumes using a Leica SP8 confocal microscope or a Prairie Ultima IV two-photon microscope. Clones composed of more than eight cells often extended the imaging volume, and therefore were counted as eight cells. The surface planar view was created by maximum-projection of the first 12-μm volume close to the surface to approximate staining results from thin sections. For co-immunostaining with GS, $580 \times 580 \times 100$-μm^3 volumes were imaged. Stitched single-plane images were processed from individual tiles by Adobe Photoshop. The number of EdU$^+$, Ki67$^+$, GS$^+$, CK19$^+$ hepatocytes was manually counted.

TRAP assays. Telomeric repeat amplification protocol (TRAP) was carried out by a previously established protocol[40]. FACS-sorted cells or homogenized tissue were lysed in NP40 buffer (25 mM HEPES-KOH, 400 mM NaCl, 1.5 mM MgCl$_2$, 10% glycerol, 0.5% NP40, and 1 mM DTT (pH 7.5) supplemented with protease inhibitors).

Ethical compliance. All animal protocols were approved by the Institutional Animal Care and Use Committee at Stanford University. All experiments complied with the relevant ethical regulations of Stanford University.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Codes are available from the corresponding author upon request.

Data availability. The source data for the RNA-seq study are available in the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE104415. Source Data for Figs. 1–5 and Extended Data Fig. 1–7, 9, 10 are available with the online version of the paper.

27. Davis, R. Cre recombinase-expressing mice generated for the NIH Neuroscience Blueprint Cre Driver Network. *MGI Direct Data Submission* http://www.credrivermice.org (2009).
28. Wu, S., Ying, G., Wu, Q. & Capecchi, M. R. A protocol for constructing gene targeting vectors: generating knockout mice for the cadherin family and beyond. *Nat. Protocols* **3**, 1056–1076 (2008).
29. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
30. Jackson, E. L. et al. Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev* **15**, 3243–3248 (2001).

31. Wu, Z., Autry, A. E., Bergan, J. F., Watabe-Uchida, M. & Dulac, C. G. Galanin neurons in the medial preoptic area govern parental behaviour. *Nature* **509**, 325–330 (2014).
32. Zolotukhin, S. et al. Recombinant adeno-associated virus purification using novel methods improves infectious titer and yield. *Gene Ther* **6**, 973–985 (1999).
33. Aurnhammer, C. et al. Universal real-time PCR for the detection and quantification of adeno-associated virus serotype 2-derived inverted terminal repeat sequences. *Hum. Gene Ther. Methods* **23**, 18–28 (2012).
34. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
35. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2009).
36. Snippert, H. J., Schepers, A. G., Delconte, G., Siersema, P. D. & Clevers, H. Slide preparation for single-cell-resolution imaging of fluorescent proteins in their three-dimensional near-native environment. *Nat. Protocols* **6**, 1221–1228 (2011).
37. Cui, W., Taub, D. D. & Gardner, K. qPrimerDepot: a primer database for quantitative real time PCR. *Nucleic Acids Res* **35**, D805–D809 (2007).
38. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
39. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
40. Kim, N. W. & Wu, F. Advances in quantification and characterization of telomerase activity by the telomeric repeat amplification protocol (TRAP). *Nucleic Acids Res* **25**, 2595–2597 (1997).
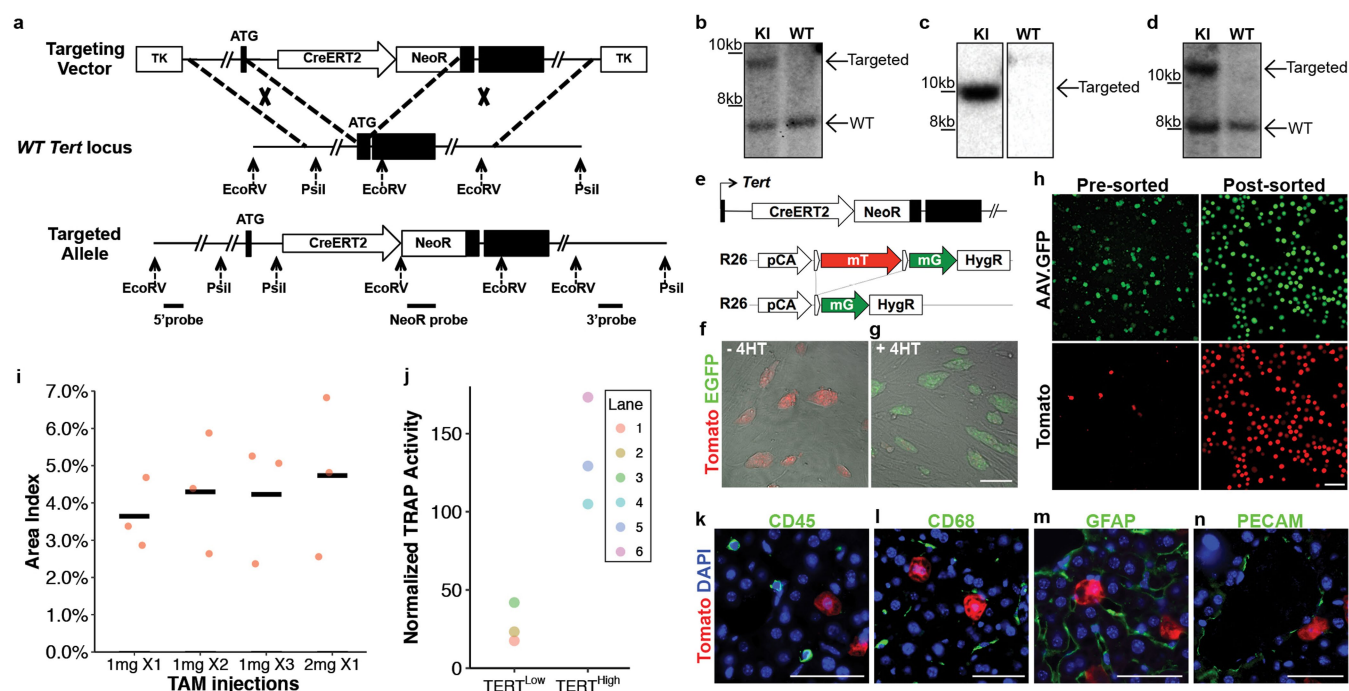
Extended Data Fig. 1 | Generation and characterization of the
Tert$^{CreERT2/+}$ **knock-in line. a**, *Tert^{CreERT2}* targeting strategy and Southern
blot strategy. **b–c**, Southern-blots using a 5′ probe (**b**), a NeoR probe (**c**),
and a 3′ probe (**d**). KI, knock-in cells; WT, wild-type cells. For gel source
data, see Supplementary Fig. 1. **e–g**, *Tert$^{CreERT2/+}$Rosa26$^{mTmG/+}$* mouse ES
cells, which respond to Cre-mediated recombination by switching from
membrane Tomato to membrane EGFP expression (**e**), showed either
membrane Tomato (**f**, overlaid on bright-field image) or membrane EGFP
(**g**, overlaid on bright-field image) in response to 500 nM 4-hydroxy
tamoxifen (4-HT). **h**, Hepatocytes from *Tert$^{CreERT2/+}$Rosa26$^{LSL-Tomato/+}$*

livers before and after FACS enrichment. **i**, Tamoxifen dose–response
curve for *Tert$^{CreERT2/+}$Rosa26$^{LSL-Tomato/+}$* livers ($n = 3$ mice for each group;
horizontal bar shows mean). **j**, Quantification of the TRAP assay shown in
Fig. 1e by densitometry. **k–n**, Co-immunofluorescence for Tomato (red)
and CD45 (**k**, blood cells, 202 cells examined), CD68 (**l**, Kupffer cells, 179
cells examined), GFAP (**m**, stellate cells, 158 cells examined) and PECAM
(**n**, endothelial cells, 167 cells examined) in *Tert$^{CreERT2/+}$Rosa26$^{LSL-Tomato/+}$*
livers after 3-day trace with DAPI (blue) staining. Experiments repeated
twice for **b–d**, **f**, **g**, **k–n**. Scale bars, 100 μm in **g**, **h**, 50 μm in **k–n**.
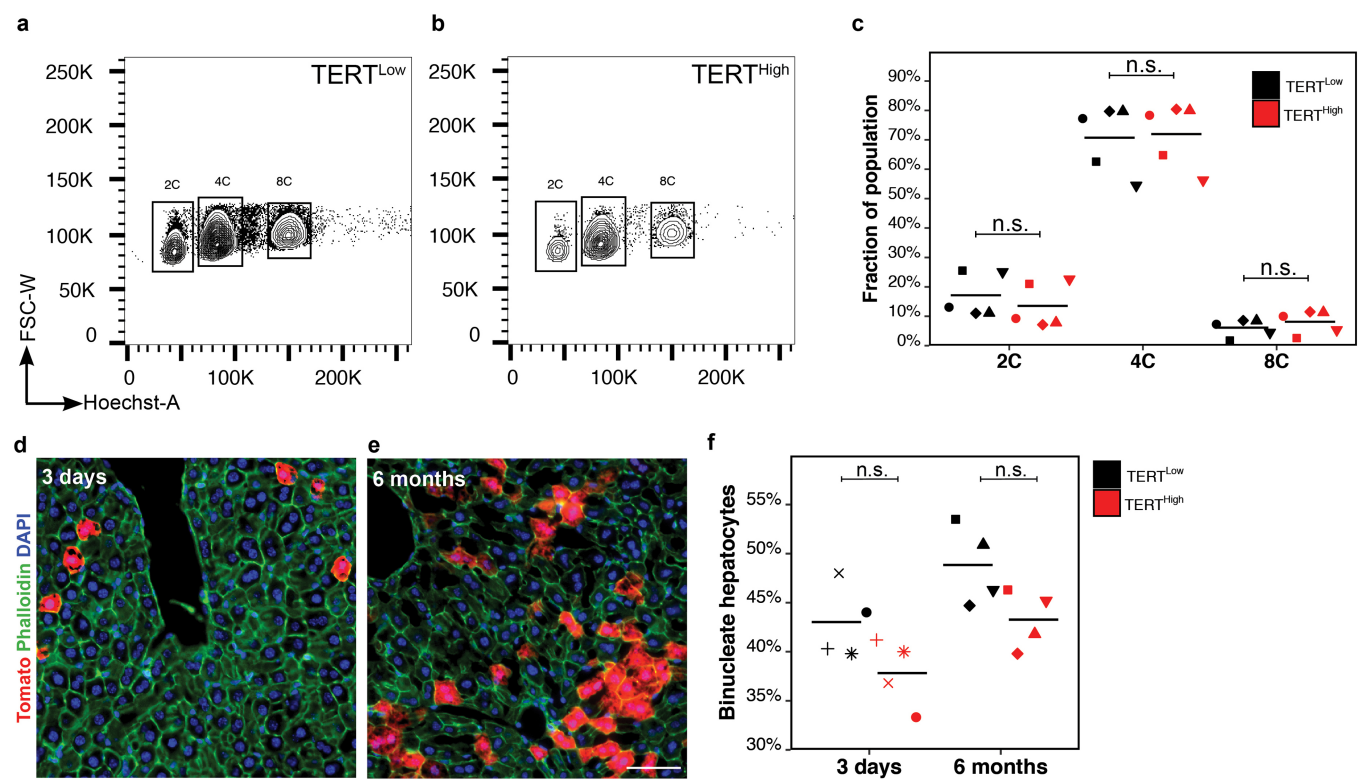
Extended Data Fig. 2 | Ploidy and nuclear profiles of the TERT^Low and TERT^High lineages. a–c, Ploidy analysis by Hoechst incorporation and FACS in TERT^Low (**a**) and TERT^High (**b**) hepatocytes. **c**, Quantification showed no significant difference between TERT^Low and TERT^High cells regarding ploidy ($n = 5$ mice, each represented by unique dot shapes).

d–f, Nucleus count by Tomato (red), phalloidin (green) and DAPI (blue) in livers traced for 3 days (**d**) and 6 months (**e**). **f**, Quantification showed no significant difference between TERT^Low and TERT^High cells in binucleus fractions ($n = 4$ mice for each group, each represented by unique dot shapes). Experiments repeated twice. Scale bar, 50 μm.
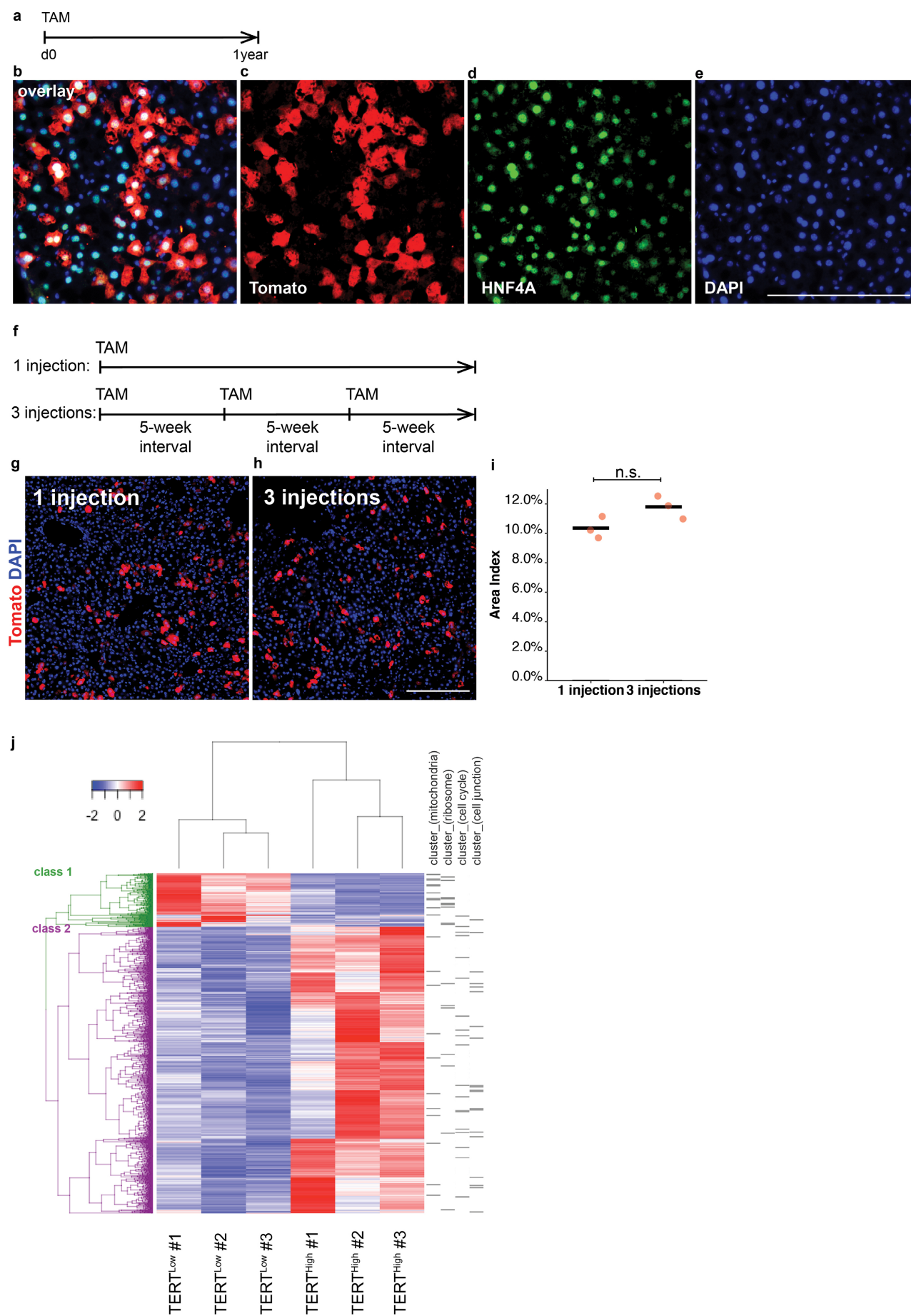
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Characterization of the lineage expansion of TERT^{High} hepatocytes. a–e, Immunofluorescence performed on *Tert$^{CreERT2/+}$Rosa26$^{LSL-Tomato/+}$* livers after one-year trace showed that only TERT^{High} hepatocytes gave rise to hepatocytes. **f–i,** Repeated injections (**f**) showed that TERT^{High} cells formed a constant proportion of the liver. Lineage expansion over one injection (**g**) and three injections (**h**) was quantified (**i**, $n = 3$ mice for each group; horizontal bars show mean). **j,** Heat map showing differentially regulated genes among all TERT^{Low} and TERT^{High} samples. Class 1 and class 2 refer to genes significantly downregulated and upregulated in TERT^{High} samples, respectively. Genes assigned to DAVID-generated annotation clusters shown on the right. Experiments repeated twice. Scale bar, 200 μm.
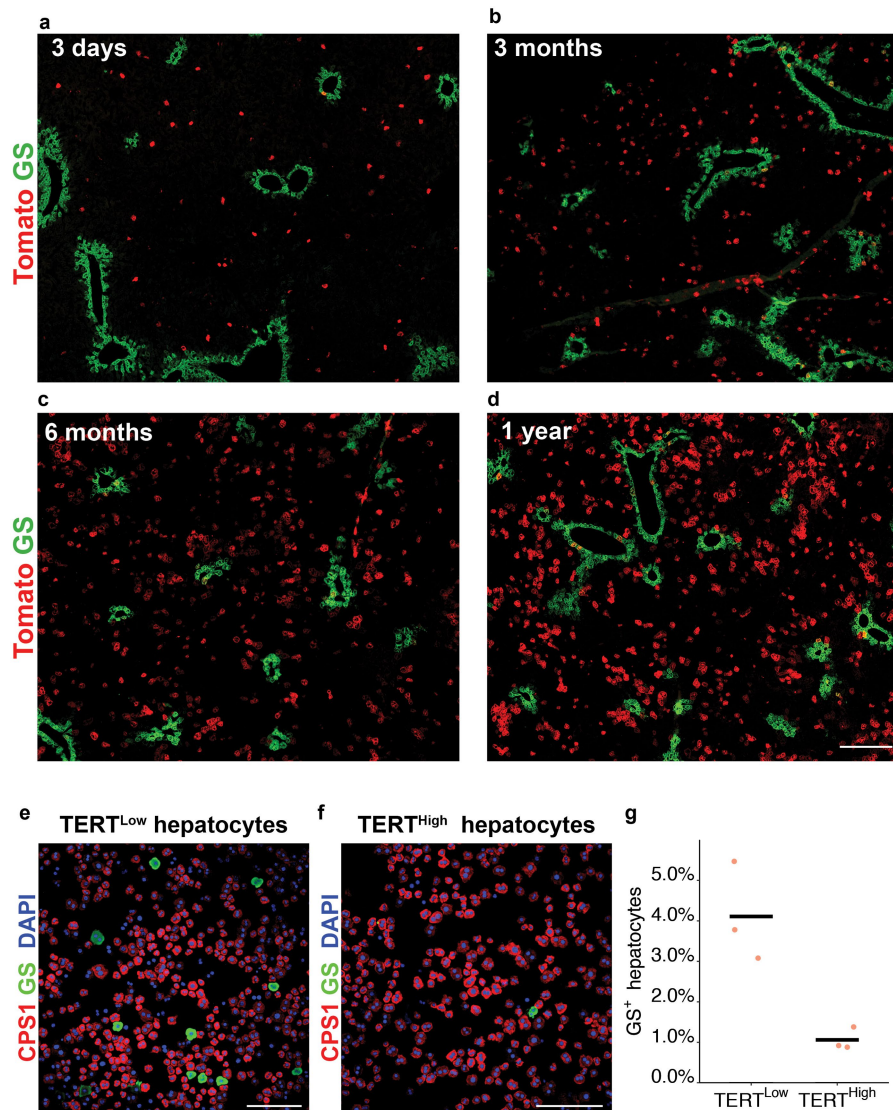
Extended Data Fig. 4 | Zonal pattern of TERT^High lineage hepatocytes.
a–d, Stitched images of immunofluorescence for Tomato protein (red) and GS (green) in liver sections from *Tert^CreERT2/+Rosa26^LSL-Tomato/+* mice treated with tamoxifen and traced for three days (**a**), three months (**b**), six months (**c**) or one year (**d**). **e–g**, FACS-isolated and cytospun hepatocytes from *Tert^CreERT2/+Rosa26^LSL-Tomato/+* mice treated with tamoxifen and traced for three days were stained for CPS1 (red) and GS (green) in TERT^Low (**e**) and TERT^High hepatocytes (**f**), and quantified for the GS^+ fraction of all cells (**g**, *n* = 3 mice; horizontal bars show mean). Experiments repeated three times. Scale bars, 200 μm.
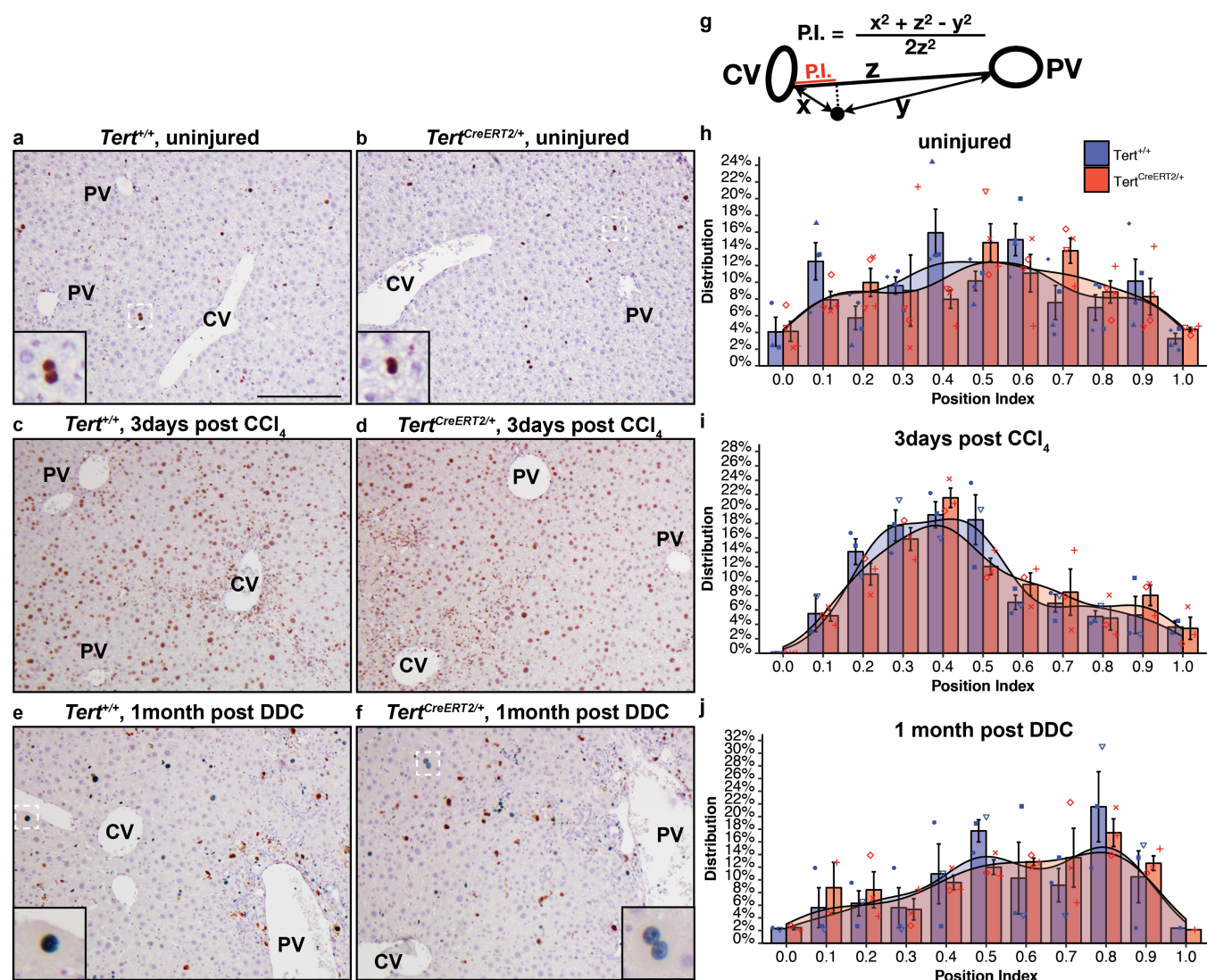
Extended Data Fig. 5 | Distribution of proliferating hepatocytes in *Tert*[+/+] and *Tert*[CreERT2/+] livers in homeostasis and after injury.
a–f, Livers were stained with anti-Ki-67 antibody by standard immunohistochemistry. **a–d**, Ki-67[+] nuclei are indicated by brown colours in uninjured livers (**a, b**), and CCl$_4$ (10 μl per 10 g weight) injured livers (**c, d**), with haematoxylin counterstain in light blue. **e, f**, Green chromogen was used to indicate Ki-67[+] nuclei in DDC (0.1%) treated livers. Hepatocyte nuclei were distinguished by size and morphology. Examples of Ki-67[+] hepatocyte nuclei are shown in insets. **g**, Quantification of Ki-67[+] hepatocytes and their distribution along the central–portal axis. The position index (P.I.) was determined by the distance to the most adjacent central vein (CV) (*x*), the distance to the most adjacent portal vein (PV) (*y*),

and the distance between the central and portal veins (*z*), following the law of cosines. **h–j**, Two-sided Kolmogorov–Smirnov tests were performed to analyse the distributions of Ki-67[+] hepatocytes along the central–portal axis. Histograms (bin-width = 0.1) and shaded curves of the kernel density estimation with Gaussian approximation are shown (mean ± s.e.m.). No significant differences were found between *Tert*[+/+] and *Tert*[CreERT2/+] livers in uninjured livers (**h**, *n* = 4 mice for each group; each mouse represented by unique dot shapes; *P* = 0.58), in CCl$_4$-injured livers (**i**, *n* = 3 mice for each group; each mouse represented by unique dot shapes, *P* = 0.32), or in DDC injured livers (**j**, *n* = 3 mice for each group; each mouse represented by unique dot shapes; *P* = 0.98). Experiments repeated twice. Scale bar, 200 μm.
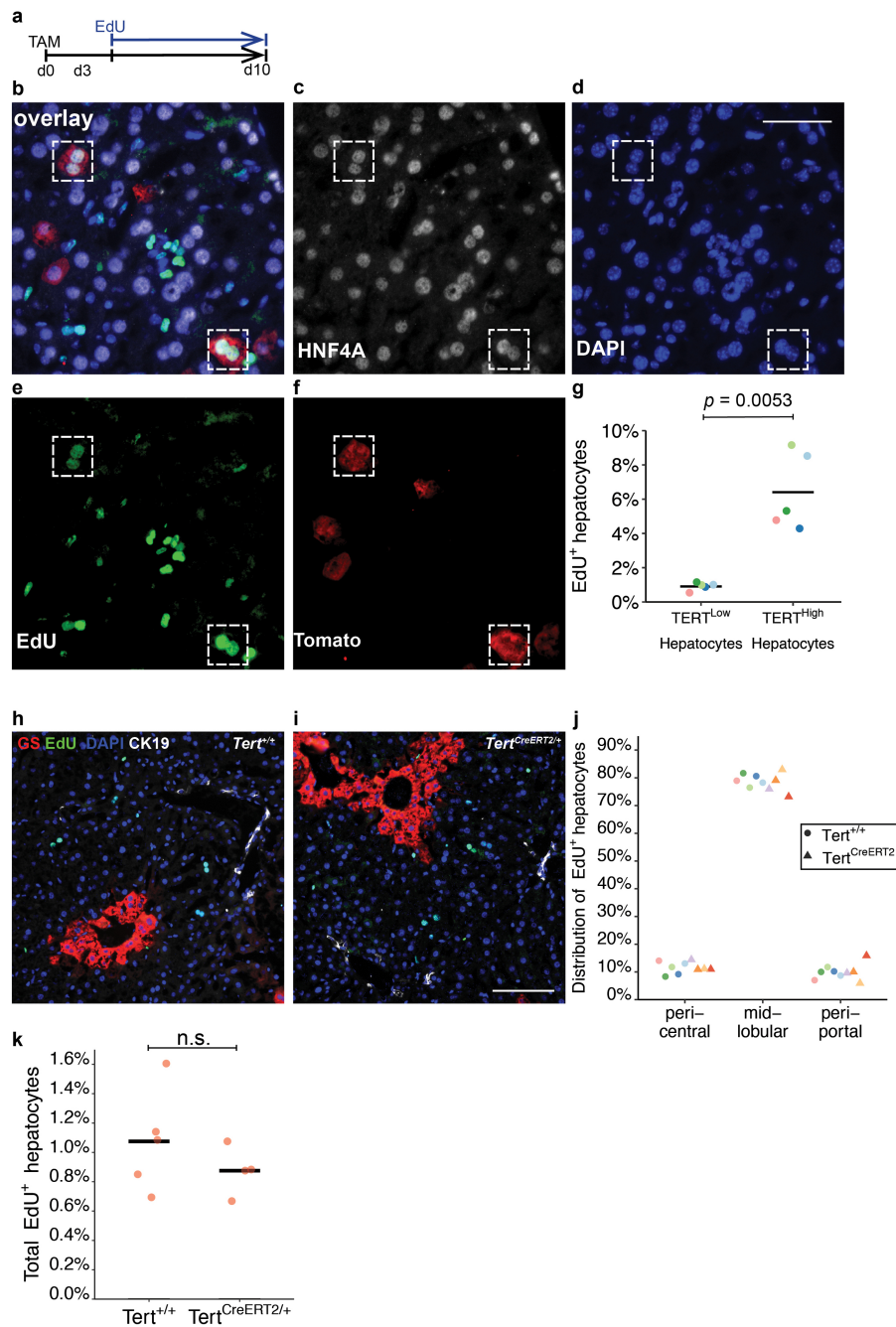
Extended Data Figure 6 | EdU incorporation assays. a, Scheme of experiments. **b–g**, EdU incorporation in livers of $Tert^{CreERT2/+}$ $Rosa26^{LSL-Tomato/+}$ mice treated with tamoxifen, traced for three days, then treated with EdU in drinking water for 7 days (1 mg ml^{-1}); overlay image (**b**), HNF4A (**c**), DAPI (**d**), EdU (**e**) and Tomato (**f**). Dashed boxes, EdU$^+$HNF4A$^+$Tomato$^+$ cells. **g**, Quantification of EdU incorporation into hepatocytes ($n = 5$ mice, each represented by unique dot colours). **h–k**, EdU incorporation into livers of $Tert^{+/+}$ (**h**) and $Tert^{CreERT2/+}$ (**i**) mice

were compared. Co-immunofluorescence for GS (red) and CK19 (white) was overlaid with EdU (green) and DAPI (blue). **j**, Quantification of the distribution of EdU$^+$ hepatocytes (pericentral, in GS$^+$ zones; periportal, 0–2 cell layers adjacent to the portal vein space or CK19$^+$ bile ducts; mid-lobular, neither pericentral nor periportal). Dot colours represent individual mice. **k**, Total EdU$^+$ hepatocytes in $Tert^{+/+}$ and $Tert^{CreERT2/+}$ livers ($n = 5$ mice for $Tert^{+/+}$ livers; $n = 4$ mice for $Tert^{CreERT2/+}$ livers). Experiments repeated twice. Scale bars, 50 μm in **d**, 200 μm in **i**.
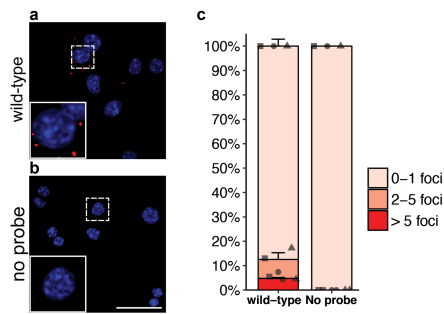
Extended Data Figure 7 | Single-molecule RNA FISH on wild-type hepatocytes. a, Experiment performed on wild-type hepatocytes isolated by FACS and cytospun. Red foci show individual *Tert* mRNA molecules. Control experiment by omitting the detection probe for *Tert*. **c**, Quantification by focus counts ($n = 3$ mice, each represented by unique dot shapes; mean + s.e.m.). Experiments repeated three times. Scale bar, $50\,\mu$m.
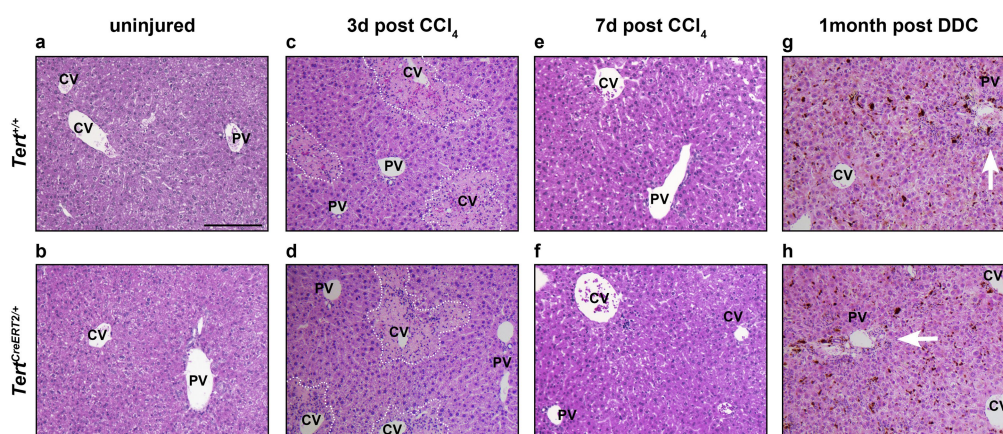
Extended Data Figure 8 | Responses of *Tert*[+/+] and *Tert*[CreERT2/+] livers to injuries. a, **b**, Haematoxylin and eosin (H&E) staining of uninjured livers. **c-d**, H&E staining of livers 3 days after CCl_4 injection. White dotted lines encircle the damaged pericentral area. **e**, **f**, H&E staining of livers 7 days after CCl_4 injection. **g**, **h**, H&E staining of livers 1 month after DDC treatment. Experiments repeated five times. Scale bar, 200 μm.
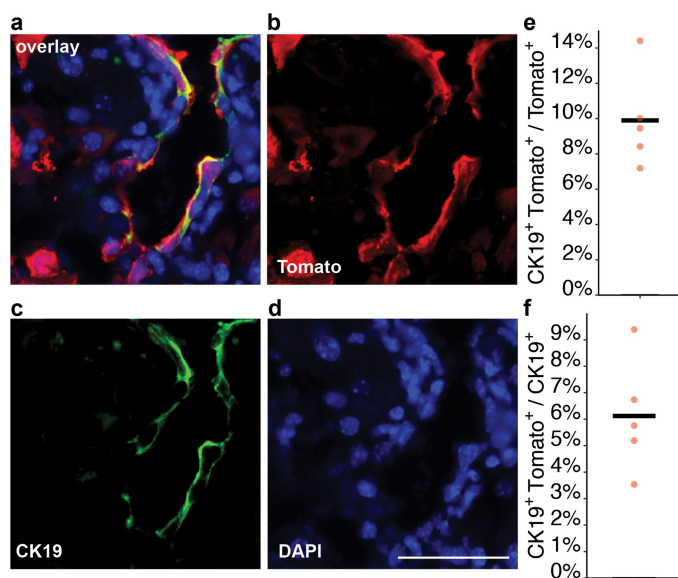
Extended Data Figure 9 | Progeny of TERT^{High} hepatocytes can adopt ductal fate after DDC injury. a–d, Immunofluorescence analysis of $Tert^{CreERT2/+}Rosa26^{LSL-Tomato/+}$ livers treated with tamoxifen and DDC, and traced for 1 month (**a**, overlay image; **b**, Tomato; **c**, CK19. **d**, DAPI). **e**, Quantification of the percentage of CK19$^+$Tomato$^+$ cells among all Tomato$^+$ cells ($n = 5$ mice, mean $\pm$ s.e.m. 10.0 $\pm$ 1.2%) **f**, Quantification of the percentage of CK19$^+$Tomato$^+$ cells among all CK19$^+$ cells ($n = 5$ mice, mean $\pm$ s.e.m. 6.1 $\pm$ 1.0%). Bars show mean. Experiments repeated three times. Scale bar, 50 μm.
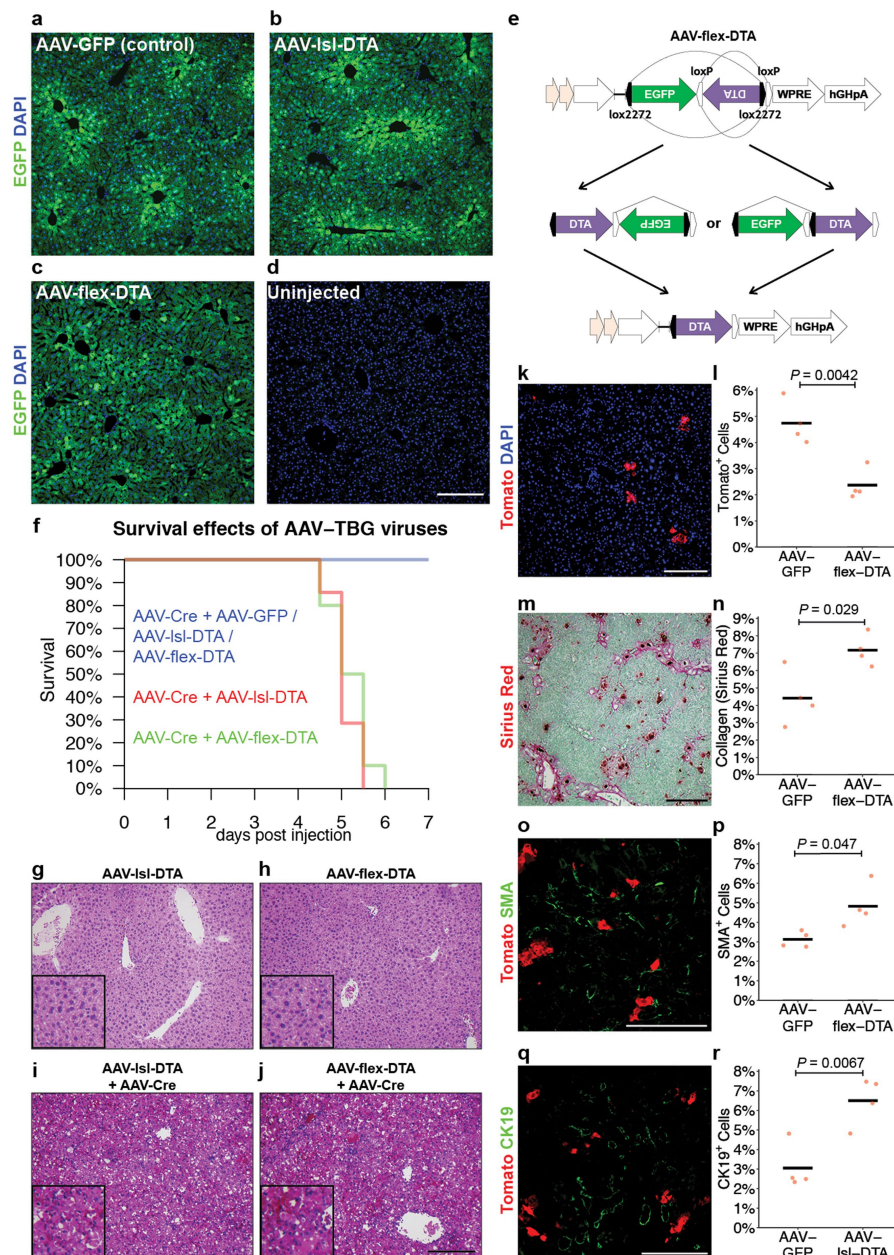
Extended Data Figure 10 | Characterization of AAV-lsl-DTA and AAV-flex-DTA. a–d, Epifluorescence of EGFP and DAPI staining of livers 4 days after injection with AAV-GFP (**a**), AAV-lsl-DTA (**b**), AAV-flex-DTA (**c**), and uninjected control (**d**). **e**, Diagram of AAV-flex-DTA and recombination events that lead to DTA expression. **f**, Survival effects of AAV-TBG viruses. Combined injection of AAV-lsl-DTA and AAV-Cre (red line) or AAV-flex-DTA and AAV-Cre (green line) lead to a narrow window of complete mortality between 4.5 and 6 days; by contrast, injection of AAV-GFP and AAV-Cre, AAV-lsl-DTA, or AAV-flex-DTA did not result in mortality. Between 4 and 6 mice were used for each regimen. Surviving mice were monitored for up to 2 months. **g–j**, H&E staining of liver sections from mice injected with AAV-lsl-DTA alone (**g**), AAV-flex-DTA alone (**h**), AAV-lsl-DTA and AAV-Cre (**i**), or AAV-flex-DTA and AAV-Cre (**j**). **k–r**, Livers injected with AAV-flex-DTA and tamoxifen showed a reduction in TERT[High] cells (**k**, **l**), as well as increases in collagen deposition (**m**, **n**), activated stellate cells (**o**, **p**) and ductal cells (**q**, **r**). Experiments repeated three times for **a–d**, and twice for **g–r**. Scale bars, 200 μm.

LETTER

Metabolic enzyme PFKFB4 activates transcriptional coactivator SRC–3 to drive breast cancer

Subhamoy Dasgupta[1,2]*, Kimal Rajapakshe[1], Bokai Zhu[1], Bryan C. Nikolai[1], Ping Yi[1], Nagireddy Putluri[1], Jong Min Choi[1], Sung Y. Jung[3], Cristian Coarfa[1], Thomas F. Westbrook[3], Xiang H.–F. Zhang[1], Charles E. Foulds[1,4], Sophia Y. Tsai[1], Ming–Jer Tsai[1] & Bert W. O'Malley[1]*

Alterations in both cell metabolism and transcriptional programs are hallmarks of cancer that sustain rapid proliferation and metastasis[1]. However, the mechanisms that control the interaction between metabolic reprogramming and transcriptional regulation remain unclear. Here we show that the metabolic enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 4 (PFKFB4) regulates transcriptional reprogramming by activating the oncogenic steroid receptor coactivator-3 (SRC-3). We used a kinome-wide RNA interference-based screening method to identify potential kinases that modulate the intrinsic SRC-3 transcriptional response. PFKFB4, a regulatory enzyme that synthesizes a potent stimulator of glycolysis[2], is found to be a robust stimulator of SRC-3 that coregulates oestrogen receptor. PFKFB4 phosphorylates SRC-3 at serine 857 and enhances its transcriptional activity, whereas either suppression of PFKFB4 or ectopic expression of a phosphorylation-deficient Ser857Ala mutant SRC-3 abolishes the SRC-3-mediated transcriptional output. Functionally, PFKFB4-driven SRC-3 activation drives glucose flux towards the pentose phosphate pathway and enables purine synthesis by transcriptionally upregulating the expression of the enzyme transketolase. In addition, the two enzymes adenosine monophosphate deaminase-1 (AMPD1) and xanthine dehydrogenase (XDH), which are involved in purine metabolism, were identified as SRC-3 targets that may or may not be directly involved in purine synthesis. Mechanistically, phosphorylation of SRC-3 at Ser857 increases its interaction with the transcription factor ATF4 by stabilizing the recruitment of SRC-3 and ATF4 to target gene promoters. Ablation of SRC-3 or PFKFB4 suppresses breast tumour growth in mice and prevents metastasis to the lung from an orthotopic setting, as does Ser857Ala-mutant SRC-3. PFKFB4 and phosphorylated SRC-3 levels are increased and correlate in oestrogen receptor-positive tumours, whereas, in patients with the basal subtype, PFKFB4 and SRC-3 drive a common protein signature that correlates with the poor survival of patients with breast cancer. These findings suggest that the Warburg pathway enzyme PFKFB4 acts as a molecular fulcrum that couples sugar metabolism to transcriptional activation by stimulating SRC-3 to promote aggressive metastatic tumours.

Among the landscape of genetic alterations that drive aggressive metastatic tumours, the transcriptional coregulator SRC-3 is one of the abundantly deregulated oncogenes[3–5]. Importantly, dynamic interactions between SRC-3 and its subsequent recruitment to target genes are delicately regulated by post-translational modifications on SRC-3[6]. Phosphorylation of SRC-3 can alter its transcriptional activity, protein stability and subcellular localization[7–9], and deregulated kinase signalling hyperactivating SRC-3 is a hallmark of many tumours[10,11]. As a starting point for identifying kinases that modulate SRC-3 transcriptional activity, we performed an unbiased RNA interference (RNAi) screening assay using a kinome library containing short interfering RNAs (siRNAs) that target 636 human kinases (median 3 siRNAs

per kinase) in the presence of a GAL4-DNA binding domain-fused SRC-3 (pBIND-SRC-3)[12] and GAL4 DNA-binding sites containing the luciferase reporter gene (pG5-luc) (Fig. 1a). The concentration of the pBIND-SRC-3 construct needed to obtain luciferase readings in a linear range was standardized along with the dose of kinase siRNAs to observe significant alterations in SRC-3 intrinsic activity (Extended Data Fig. 1a, b). As a positive control, we used siRNAs that target PRKCZ1, a protein kinase known to activate SRC-3[13], and compared the repression of the coregulator activity after kinase knockdown with non-targeting control green fluorescent protein (GFP) siRNAs (Extended Data Fig. 1c). Kinome-wide screening identified several kinases as modulators of SRC-3 activity (Fig. 1b, Extended Data Fig. 1d and Supplementary Table 1), as either stimulators or repressors compared to the controls (Extended Data Fig. 1e).

Ten kinases were designated as reproducible and significant hits in the screen (Fig. 1c and Extended Data Fig. 1f), among which metabolic kinase PFKFB4 was identified as the most robust positive regulator of SRC-3 activity. A secondary screen coupled with growth assays to identify the top-hit kinases that drive cancer cell proliferation also identified PFKFB4 as the most dominant kinase that regulates cellular proliferation (Extended Data Fig. 1g). Silencing of PFKFB4 with different short hairpin RNAs (shRNAs) and siRNAs decreased SRC-3 activity (Extended Data Fig. 2a, b) in several cancer lines with reduced PFKFB4 levels (Extended Data Fig. 2c, d), whereas ectopic overexpression of PFKFB4 using adenoviral infection (ad-PFKFB4) enhanced SRC-3 activity (Fig. 1d). Interestingly, SRC-3 protein levels were increased after ectopic PFKFB4 expression (Fig. 1e), but *SRC-3* (also known as *NCOA3*) mRNA levels were not affected (Extended Data Fig. 2e), and proximity ligation assays support a direct interaction between SRC-3 and PFKFB4, consistent with PFKFB4-dependent regulation of SRC-3 activity (Extended Data Fig. 2f).

PFKFB4 is a bifunctional metabolic enzyme that synthesizes fructose 2,6-bisphosphate (F2,6-BP), an important sugar-phosphate metabolite that stimulates glycolysis[14]. PFKFB4 dovetails two antagonistic properties involving a kinase reaction synthesizing F2,6-BP from fructose-6-phosphate (F6P) and ATP, and conversely hydrolysing F2,6-BP into F6P and inorganic phosphate (P_i) via its phosphatase activity[15,16]. These properties of PFKFB4 prompted us to investigate whether PFKFB4-catalysed enzymatic reactions could increase phosphorylation on SRC-3. An in vitro enzymatic reaction containing F6P, ATP and varying concentrations of recombinant PFKFB4 enzyme were incubated with purified full-length SRC-3 protein. Increasing the amount of PFKFB4 enzyme in the reaction concomitantly enhanced the Ser/Thr phosphorylation of SRC-3, indicating that the metabolic enzyme PFKFB4 can phosphorylate a protein substrate (Extended Data Fig. 3a). We investigated the phosphate donor in the PFKFB4 kinase reaction, and identified ATP as being required for SRC-3 phosphorylation by PFKFB4 (Extended Data Fig. 3b). These findings suggest that PFKFB4 can function as a protein kinase to phosphorylate SRC-3 by transferring

[1]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [2]Department of Cell Stress Biology, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. [3]Verna & Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, USA. [4]Center for Precision Environmental Health, Baylor College of Medicine, Houston, TX, USA. *e-mail: subhamoy.dasgupta@roswellpark.org; berto@bcm.edu
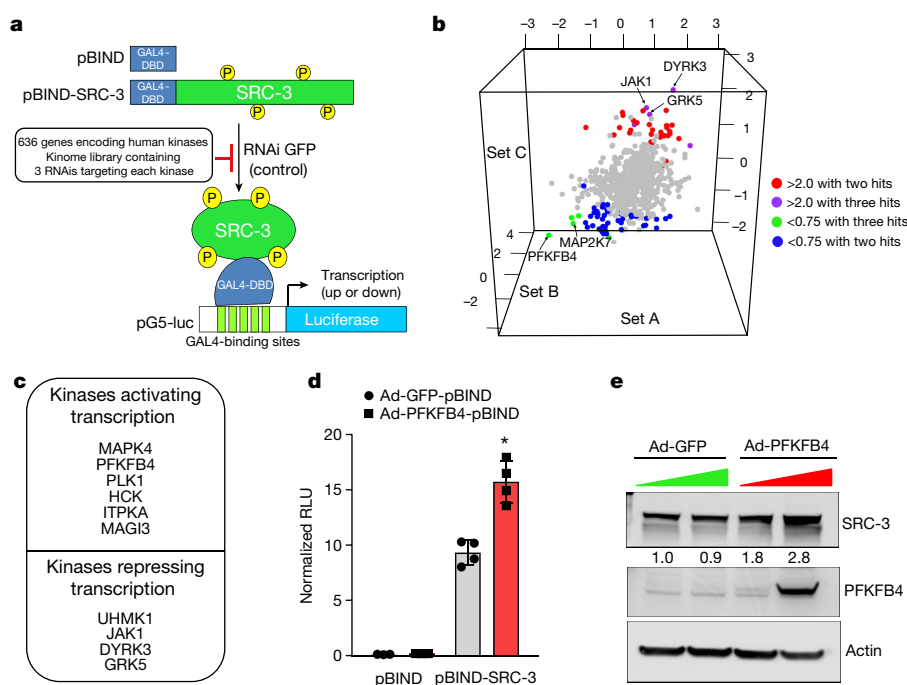
Fig. 1 | PFKFB4 is an essential activator of transcriptional coregulator SRC-3. a, Schematics showing the RNAi kinome library screening with SRC-3 transcriptional activity assay using the GAL4 DNA-binding site luciferase reporter (pG5-luc) along with the GAL4 DNA-binding domain (DBD) full-length SRC-3 fusion (pBIND-SRC-3) construct or the control pBIND construct as readout. **b**, $\log_2$ fold change in SRC-3 activity with three siRNAs per kinase represented as sets A, B and C in the 3D plot ($n = 3$ biologically independent samples targeted by siRNAs). Suppression of kinases either increased SRC-3 activity, with 2 out of 3 siRNAs (red) or 3 out of 3 siRNAs (purple), or reduced SRC-3 activity, with 3 out of 3 siRNAs (green) or 2 out of 3 (blue). **c**, Kinases scoring reproducibly in modulating SRC-3 activity. **d**, SRC-3 activity in MCF-7 cells transduced with adenovirus expressing PFKFB4 or GFP (ad-PFKFB4 or ad-GFP) and co-transfected with pBIND ($n = 3$) or pBIND-SRC-3 ($n = 4$). RLU, relative luciferase units. Data are mean ± s.d. *$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test. **e**, Protein expression of SRC-3, PFKFB4 and actin in MCF-7 cells transduced with adenovirus expressing PFKFB4 or GFP control. See Source Data for exact P values; n represents biologically independent samples.

a phosphate group from ATP. To confirm this observation, we performed a kinase assay using [γ-^{32}P]ATP as the phosphate donor and observed enhanced incorporation of phosphate from [γ-^{32}P]ATP into SRC-3 protein upon increasing concentrations of the PFKFB4 kinase in the reaction (Fig. 2a). To identify the phosphorylation site(s) on SRC-3, we used recombinant glutathione S-transferase (GST)-fused SRC-3 fragments encoding various domains (Extended Data Fig. 3c) as substrates for an in vitro kinase reaction, and found that only the CBP-interacting domain (CID) of SRC-3[17] is phosphorylated by PFKFB4 (Fig. 2b). In vitro phosphorylated GST–SRC-3-CID protein was then analysed by mass spectrometry, and only one serine residue (Ser857) was identified as a phosphorylation target of PFKFB4 (Extended Data Fig. 3d). Consistent with this identification, mutation of Ser857 to alanine (Ser857Ala) abolished the phosphorylation of SRC-3-CID by PFKFB4 in vitro (Extended Data Fig. 3e), confirming that PFKFB4 phosphorylates oncogenic coregulator SRC-3 at Ser857.

Because increased glucose metabolism stimulates the kinase activity of PFKFB4 required to maintain steady glycolysis[18], we measured the levels of phosphorylated SRC-3 (pSRC-3) under these conditions. HEK293T cells were transfected with Flag-tagged SRC-3 and PFKFB4, and then stimulated with an increasing concentration of glucose in culture medium, which revealed enhanced phosphorylation of SRC-3 (Fig. 2c). Next we investigated the levels of pSRC-3-S857 in breast cancer cells under conditions of active glycolysis by immunoblotting with a pSRC-3-Ser857-antibody. MDA-MB-231 cells growing under a normal glucose condition (25 mM) showed robust phosphorylation of SRC-3 at Ser857 compared to tumour cells cultured in low glucose conditions (5 mM) (Fig. 2d). Withdrawing glucose from the medium after growth in normal glucose conditions (25 mM) resulted in significant loss of SRC-3 phosphorylation (Fig. 2d). Moreover, stable knock-down of PFKFB4 using two different shRNA constructs (shPFK#09 and shPFK#20) (Fig. 2d and Extended Data Fig. 3f) abolished pSRC-3-Ser857 levels in breast cancer cells cultured in 25 mM glucose, indicating that PFKFB4-dependent SRC-3 phosphorylation on Ser857 is a highly selective modification under conditions conducive to active glycolysis. We expressed the phosphorylation-defective mutant SRC-3(Ser857Ala) or wild-type SRC-3 protein in SRC-3-ablated cells, and under conditions of active glycolysis the levels of pSRC-3-Ser857 are increased in wild-type SRC-3 cells compared to the SRC-3(Ser857Ala) mutant cells (Extended Data Fig. 3g). Importantly, the introduction

of fructose-1,6-bisphosphate (FBP) alone into glucose-starved cells permeabilized with streptolysin O rescued pSRC-3-Ser857 levels (Extended Data Fig. 3h), indicating that this phosphorylation event is linked to the energy status of the cell[19].

To measure the importance of this modification on the intrinsic activity of SRC-3, we transduced cancer cells with adenovirus expressing PFKFB4 (ad-PFKFB4) or control GFP, and cultured the transduced cells in the presence of normal glucose (25 mM) or low glucose (5 mM) levels. Enhanced expression of PFKFB4 along with glucose stimulation significantly increased the transcriptional activity of SRC-3 (pBIND-SRC-3) compared to cells cultured in low glucose conditions, suggesting that PFKFB4-dependent SRC-3 phosphorylation is important for the coactivator-driven transcriptional response (Extended Data Fig. 3i). To substantiate this observation, we used the phosphorylation-deficient pBIND-SRC-3(Ser857Ala) mutant or the phosphorylation-mimic pBIND-SRC-3(Ser857Glu) mutant in a similar transcriptional activation assay and found that the Ser857Ala mutant was significantly refractory to glucose-dependent PFKFB4 signalling (Fig. 2e). The Ser857Glu mutant was constitutively active even at low levels of glucose, and glucose stimulation failed to show any further activation (Extended Data Fig. 4a). Previous studies have identified several crucial sites in the kinase domain of PFKFB4 that are important for ATP binding[14]. When mutated to alanine, residues Gly46, Pro48, Gly51, Arg229 and Arg237 significantly decrease the binding affinity for ATP and result in reduced PFKFB4 kinase activity. We expressed these mutants in PFKFB4-silenced breast cancer cells and transcriptional assays confirmed significantly reduced SRC-3 activity and Ser857 phosphorylation (Extended Data Fig. 4b, c). Because SRC-3 is an established oestrogen receptor (ER) coactivator, we investigated the importance of glucose-dependent PFKFB4 signalling on ER-mediated transcriptional activity. MCF-7 cells stably expressing an oestradiol (E2)-ER-dependent luciferase reporter gene (ERE-MAR-Luc cells)[20] were used to assay ER activity as a function of E2 and glucose in the medium. Glucose addition enhanced ER activity, whereas low glucose or SRC-3 silencing significantly repressed transcriptional output in response to E2 (Extended Data Fig. 4d). Overexpression of PFKFB4 enhanced ER activity only in cells treated with E2 and glucose, whereas this PFKFB4-dependent increase in ER activity is repressed upon SRC-3 ablation (Extended Data Fig. 4e). Consistent with this observation, the SRC-3(Ser857Ala) mutant failed to rescue the growth of
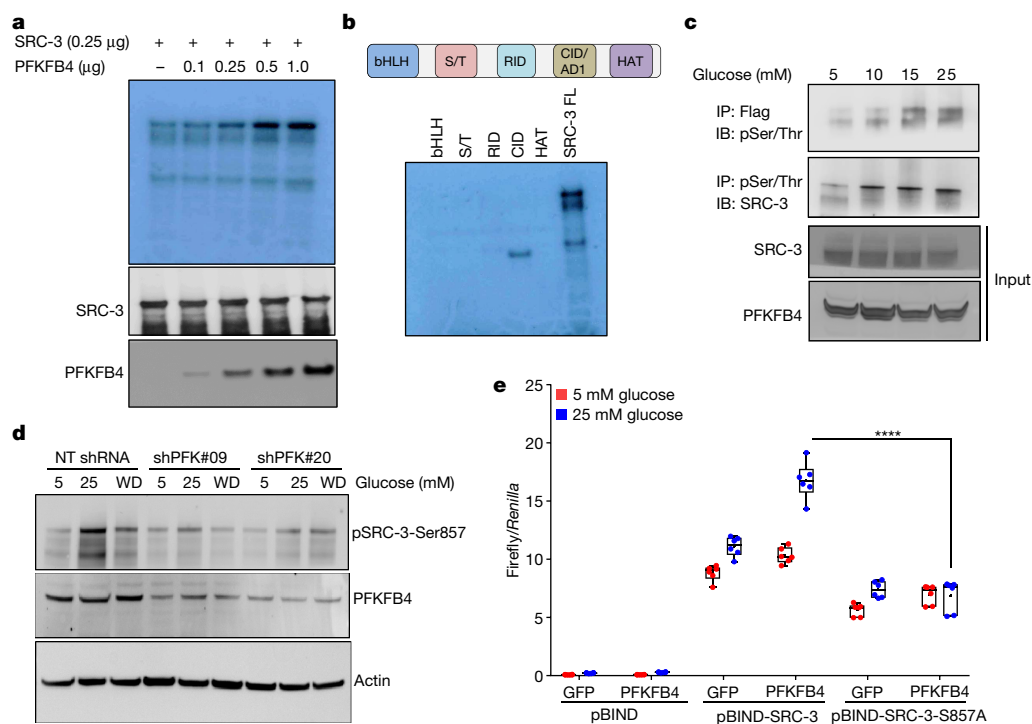
Fig. 2 | PFKFB4 phosphorylates SRC-3 by functioning as a protein kinase. a, Top, recombinant GST-fused PFKFB4 incubated with full-length SRC-3 in the presence of [^{32}P]ATP in an in vitro kinase assay. Bottom, SRC-3 and PFKFB4 protein levels were analysed by immunoblotting. **b**, In vitro kinase assay of PFKFB4 in the presence of SRC-3 fragments expressing different domains or full-length (FL) SRC-3. **c**, HEK293T cells expressing Flag-tagged SRC-3 and PFKFB4 cultured in different concentrations of glucose and immunoprecipitated by Flag or pSer/Thr antibodies followed by immunoblotting. **d**, MDA-MB-231 cells stably expressing shRNAs targeting PFKFB4 (shPFK#09 and shPFK#20) or control non-targeting (NT) shRNA grown in the presence of 5 mM or 25 mM glucose, or after glucose withdrawl (WD), in which cells were cultured in 25 mM glucose for 24 h and then switched to 5 mM glucose for 6 h. Protein levels of pSRC-3-Ser857, PFKFB4 and β-actin were detected by immunoblotting. **e**, HEK293T cells expressing pBIND, pBIND-SRC-3 or pBIND-SRC-3(Ser857Ala) were transduced with adenoviruses expressing GFP or PFKFB4, and cultured in 5 mM or 25 mM glucose followed by luciferase assay. Boxes represent the twenty-fifth and seventy-fifth percentiles, lines represent median, whiskers showing minimum and maximum points, and plus symbol indicates the mean. $n = 6$ biologically independent experiments. ****$P < 0.000001$, two-way ANOVA with Tukey's multiple comparisons test. Data in **a**–**e** are representative of three biologically independent experiments with similar results. See Source Data for exact P values.

SRC-3-depleted cells compared to wild-type SRC-3 (Extended Data Fig. 4f). These findings suggest that in glycolytic breast tumours, PFKFB4 and SRC-3 can also hyperactivate ER activity in the presence of E2, and phosphorylation of SRC-3 at Ser857 is a critical mark required for transcriptional responses.

PFKFB4 is an important regulator of glucose metabolism and directs metabolic pathways required for biosynthesis of macromolecules to sustain rapid proliferation in cancer cells[2]. To identify the physiological role of PFKFB4-dependent SRC-3 activation in tumour metabolism, we performed an unbiased phenotypic screen to identify the metabolites that are preferentially used by SRC-3-overexpressing cells. For this we used a phenotype microarray analysis[21] containing 93 metabolites (Supplementary Table 2) arrayed in a microplate and measured in real-time the importance of these metabolites in supporting SRC-3-dependent growth. We transduced mammary epithelial MCF10A cells (with relatively low endogenous SRC-3) with adenovirus expressing GFP or SRC-3 followed by the phenotype screen for 24 h. We identified enhanced proliferation of cells with gain-in SRC-3 expression under conditions of glucose and purines such as adenosine and inosine (Extended Data Fig. 5a–d). To investigate the role of SRC-3 further and determine how activation by PFKFB4 affects its regulation of metabolism in breast cancer cells, we performed mass spectrometry-based metabolic profiling of MDA-MB-231 cells expressing shRNAs that target *SRC-3* or *PFKFB4*. Ablation of either SRC-3 or PFKFB4 significantly reduced the intracellular pools of ribose-5P (R5P), and purine nucleotides and intermediates, such as adenosine, xanthine and guanine (Fig. 3a and Extended Data Fig. 5e, f). Overexpression of SRC-3 in MCF10A cells also confirmed increased pools of purines (Extended Data Fig. 5g). To measure the

direct contribution of PFKFB4 and SRC-3 regulation of glucose flux towards the pentose phosphate pathway (PPP), we used isotope-labelled [6-^{13}C]glucose to trace the carbon flow[22]. PFKFB4 and SRC-3 depletion significantly reduced the ^{13}C-enrichment of ribulose-5P/xylulose-5P, important intermediary metabolites in the PPP and rate-limiting precursors for purine biosynthesis (Extended Data Fig. 6a). We investigated whether exogenous addition of purines could rescue the reduced growth rate of SRC-3-deficient breast cancer cells. As expected, loss of SRC-3 suppressed the growth of MDA-MB-231 and MCF-7 breast cancer cells[23], whereas supplementation of purines in the culture medium significantly rescued the growth defect, indicating that SRC-3 expression is crucial for the synthesis of purines for growth (Fig. 3b).

To identify the potential mechanisms of apparent SRC-3-driven purine synthesis, we performed gene expression analysis of enzymes involved in the PPP and purine synthesis. Knockdown of SRC-3 reduced the mRNA expression of transketolase (*TKT*), adenosine monophosphate deaminase 1 (*AMPD1*), and xanthine dehydrogenase (*XDH*) (Extended Data Fig. 6b, c). These SRC-3 target genes were also found to be regulated by PFKFB4 knockdown (Fig. 3c) and their expression was significantly enhanced in actively glycolytic breast cancer cells (Extended Data Fig. 6d). TKT is a major enzyme mediating non-oxidative PPP, whereas XDH and AMPD1 traditionally known to regulate purine catabolism are found to be regulated by SRC-3[24,25]. Whether the switch in roles by these reversible enzymes XDH and AMPD1 depend on tumour metabolic state needs further investigation. Similarly, [6-^{13}C]glucose isotope-tracing experiments also confirmed reduced levels of TKT products seduheptulose-7P (S7P) and erythrose-4P (E4P) upon PFKFB4 or SRC-3 knockdown (Extended Data Fig. 6e, f).
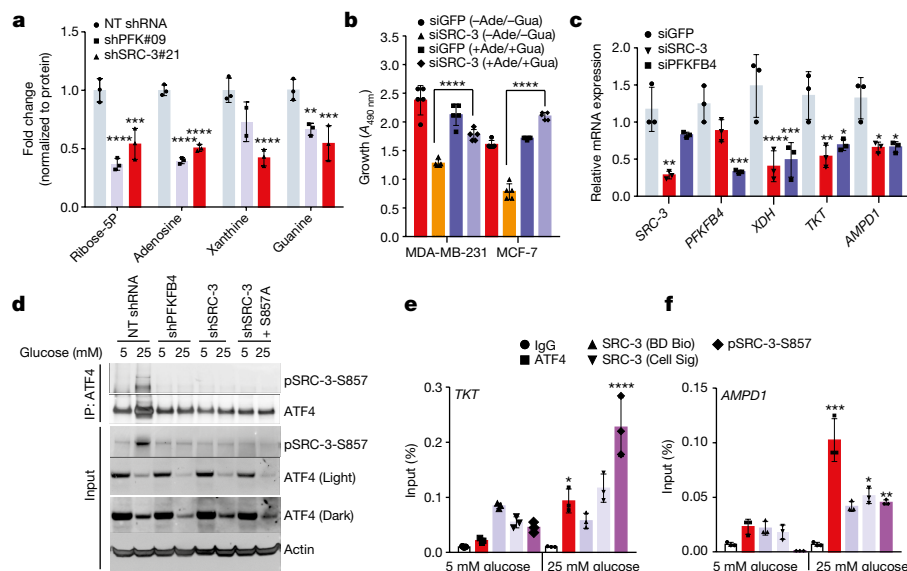
Fig. 3 | SRC-3 phosphorylation by PFKFB4 enhances gene expression of metabolic enzymes. a, Relative levels of metabolites altered by shRNAs against *PFKFB4* or *SRC-3* compared to control non-targeting shRNA in MDA-MB231 cells. $n = 3$ biologically independent samples. **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, two-way ANOVA with Tukey's multiple comparisons test. **b**, Relative proliferation of MDA-MB-231 and MCF-7 cells 4 days after treatment with siRNA targeting GFP (control) or *SRC-3* under the conditions indicated. Ade, adenine; Gua, guanine. $n = 5$ biologically independent replicates. ****$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test. **c**, mRNA expression of metabolic enzymes *TKT*, *XDH* and *AMPD1* in MDA-MB-231 cells after treatment with siRNAs targeting GFP (control), *PFKFB4* or *SRC-3*. $n = 3$ biologically independent samples. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test. **d**, Immunoprecipitation (IP) of ATF4 from MDA-MB-231 cells

grown in 5 mM or 25 mM glucose after treatment with *PFKFB4* shRNA, *SRC-3* shRNA, or non-targeting control shRNA, or after *SRC-3* shRNA plus re-expression of SRC-3(Ser857Ala). Levels of pSRC-3-Ser857 associated with ATF4 were detected by immunoblotting. IgG light chain conjugated to horseradish peroxidase (HRP) was used to probe ATF4 in immunoblotting. The ATF4 blot was exposed for shorter (light) and longer (dark) time points to visualize faint bands. **e, f**, ChIP of ATF4, total SRC-3, and pSRC-3-S857 followed by qPCR from MDA-MB-231 cells treated with 5 mM or 25 mM glucose compared to an IgG isotype control. SRC-3 antibodies were either from BD Biosciences or Cell Signaling Technology. *TKT* (**e**) and *AMPD1* (**f**) expression are shown. $n = 3$ biologically independent samples used for ChIP. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test compared to 5 mM glucose groups. See Source Data for exact P values. Unless stated otherwise, data are mean ± s.d.

To confirm that these genes are direct targets of SRC-3, we re-expressed SRC-3 in MDA-MB-231 cells with depleted levels of endogenous SRC-3 protein (Extended Data Fig. 3g) and observed significant restoration of SRC-3 target genes (Extended Data Fig. 7a). The addition of exogenous purines also restored the primary growth defects in PFKFB4-silenced MDA-MB-231 cells (Extended Data Fig. 7b), with a decreased incorporation of [U-^{13}C]glucose carbon into purines (Extended Data Fig. 7c). Although the metabolic effects may or may not be directly regulated by target genes AMPD1 or XDH, our findings indicate that PFKFB4 and SRC-3 mutually cooperate to drive glucose flux towards purine generation.

To define how PFKFB4 phosphorylation of SRC-3 affects transcriptional regulation of the three commonly regulated purine biosynthesis genes defined above, we analysed the chromatin occupancy of SRC-3 on the promoters of *TKT*, *XDH* and *AMPD1* using an existing in silico analysis of SRC-3 chromatin immunoprecipitation followed by sequencing (ChIP–seq) dataset[26]. We identified strong overlap of SRC-3 occupancy with activating transcription factor 4 (ATF4)-binding sites[27] on the three target genes (Extended Data Fig. 8a, b). Interestingly, ATF4 has been recently identified to promote purine synthesis in response to growth signals[28]. To validate whether SRC-3 interacts with ATF4, we immunoprecipitated ATF4 from MDA-MB-231 cells growing in either 25 mM or 5 mM glucose. Under conditions of enhanced glycolysis, the interaction of ATF4 with pSRC-3-Ser857 increased robustly although the total ATF4 protein level was lower owing to reduced nutrient stress compared to 5 mM glucose treatment. However, the loss of PFKFB4, SRC-3, or re-expression of SRC-3(Ser857Ala) in SRC-3-knockdown cancer cells greatly reduced the association (Fig. 3d). Next we performed ChIP and quantitative PCR (ChIP–qPCR) to measure the chromatin occupancy of ATF4, pSRC-3-Ser857 and SRC-3 on the target gene promoters. Breast cancer cells growing in the presence of

25 mM glucose showed increased occupancy of ATF4 and pSRC-3-Ser857 on *TKT* (Fig. 3e), *XDH* (Extended Data Fig. 8c) and *AMPD1* (Fig. 3f) promoters, whereas the loss of SRC-3 or PFKFB4 significantly reduced ATF4 chromatin occupancy on *AMPD1* (Extended Data Fig. 8d). In addition, we found SRC-3 recruitment to the gene promoters is dependent on ATF4, as knockdown of ATF4 significantly reduced target gene expression and pSRC-3-Ser857 promoter occupancy (Extended Data Fig. 8e, f). These findings demonstrate that in actively glycolytic breast cancers, PFKFB4-dependent phosphorylation of SRC-3 at Ser857 promotes interaction with the transcription factor ATF4, thereby stabilizing the complex on chromatin and driving transcription of key metabolic enzymes.

To study whether suppression of PFKFB4 or SRC-3 can affect the growth of breast tumours in vivo (Fig. 4a), we implanted MDA-MB-231 cells stably expressing non-targeting shRNA, *SRC-3* shRNA (Extended Data Fig. 3g) and *PFKFB4* shRNA (Fig. 2d) into the mammary fat pad of female nude mice. Compared to control mice, genetic loss of SRC-3 or PFKFB4 exhibited substantially reduced tumour growth and volume (Fig. 4b and Extended Data Fig. 9a). Immunostaining with a human Ki67 antibody showed significantly reduced proliferative cells in SRC-3- or PFKFB4-ablated tumours compared to controls (Extended Data Fig. 9b, c). To evaluate the functional significance of the Ser857 phosphorylation of SRC-3 in breast tumour progression, we stably expressed shRNA-resistant wild type or SRC-3(Ser857Ala) in MDA-MB-231 cells with suppressed expression of endogenous SRC-3 protein (Extended Data Fig. 3g). Rescuing expression with the exogenous wild-type SRC-3 construct in SRC-3-depleted cells completely restored the growth of the breast tumours (Extended Data Fig. 9d), whereas the phosphorylation-deficient Ser857Ala mutant (Extended Data Fig. 9d, e) was partially resistant to tumorigenesis six weeks after grafting the tumour cells (Fig. 4b and Extended Data Fig. 9a).
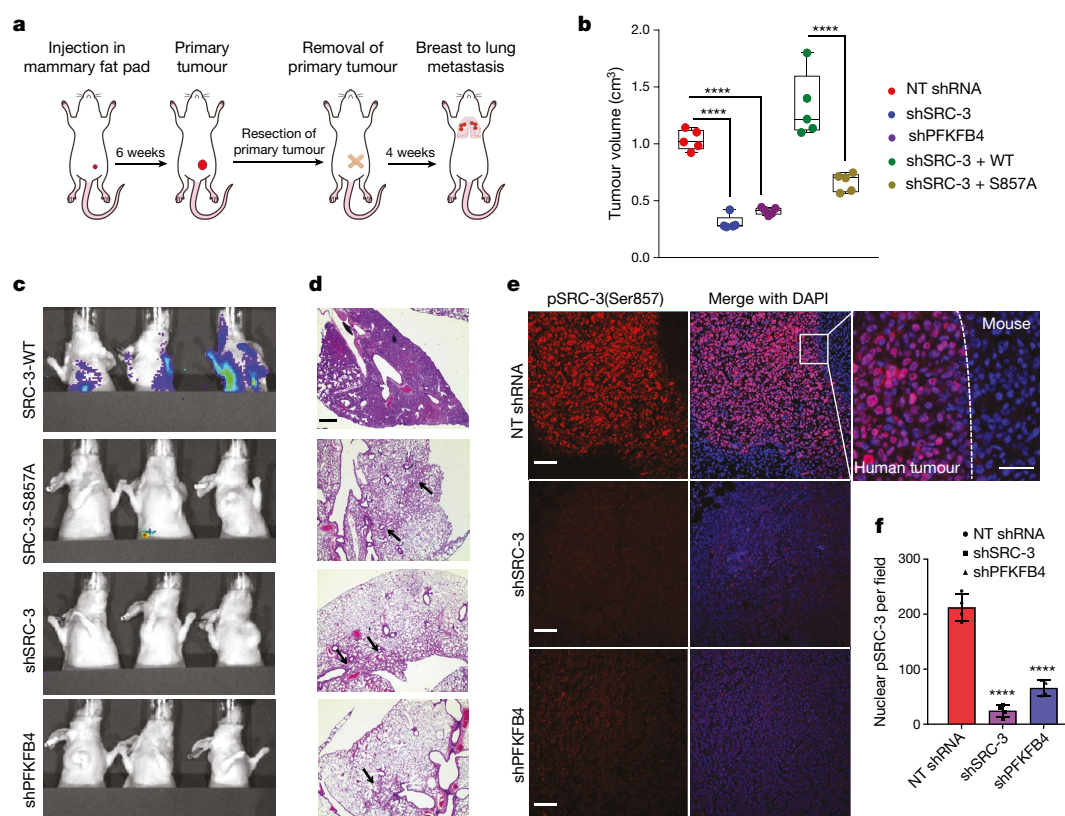
Fig. 4 | Activation of the PFKFB4–SRC-3 axis drives breast tumour primary growth and metastasis. a–g, MDA-MB-231 cells stably expressing *SRC-3* shRNA, *PFKFB4* shRNA or *SRC-3* shRNA plus wild-type (WT) SRC-3 or the SRC-3(Ser857Ala) mutant were injected into nude female mice. **a**, Schematics of the in vivo orthotopic xenograft experiment. Tumour cells were injected in the mammary fat pad ($n = 5$ mice) and after 6 weeks primary tumours were resected out and animals were monitored by bioluminescence. **b**, Tumour volume. ****$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test. $n = 5$. Boxes are as in Fig. 2e. **c**, Bioluminescence imaging of animals 4 weeks after surgery. Representative images of three animals are shown from $n = 5$ mice for wild-type SRC-3, SRC(Ser857Ala) and *PFKFB4* shRNA; and $n = 4$ mice for *SRC-3* shRNA. Residual or recurrence tumours at primary

sites were masked with black paper to visualize lung lesions. **d**, Histology images showing lung sections stained with haematoxylin and eosin. Arrows indicate micro-metastasis lesions. Scale bar, 100 μm. Data shown are representative of four fields per slide from $n = 5$ animals per group. **e**, Immunohistochemical images from primary tumours demonstrating pSRC-3-Ser857 expression (red) co-stained with DAPI (blue). Scale bars, 100 μm. Magnified image in the box shows the tumour boundary as indicated by the dotted line. Scale bar, 200 μm. **f**, Quantification of nuclear-stained pSRC-3-Ser857 in each group. Average of four fields per slide from $n = 5$ mice per group. ****$P = 0.0001$, one-way ANOVA with Dunnett's multiple comparisons test. See Source Data for exact P values. Unless stated otherwise, data are mean ± s.d.

After resecting out the primary tumours, we allowed the animals to survive for four more weeks with weekly bioluminescence imaging (Fig. 4a) to evaluate metastatic potential. Animals with primary tumours expressing wild-type SRC-3 developed profound lung metastasis with morbid hunched back posture, whereas suppression of SRC-3 or PFKFB4 or expression of the SRC-3(Ser857Ala) phosphorylation-deficient mutant all showed markedly reduced lung lesions (Fig. 4c and Extended Data Fig. 9f). Pathological analysis identified only a few micro-metastatic lesions in the lungs of animals with SRC-3(Ser857Ala), or SRC-3- and PFKFB4-ablated primary tumours (Fig. 4d), with no observed health issues during the four weeks after surgery. These findings demonstrate that SRC-3 and PFKFB4 are drivers of basal-subtype breast tumour growth and that phosphorylation of SRC-3 at the Ser857 site is crucial for metastatic progression of the disease. Immunostaining of the primary tumours with a pSRC-3-Ser857 antibody detected increased nuclear-localized human SRC-3 in the tumours collected from wild-type animals that progressed to aggressive metastatic disease, whereas PFKFB4- or SRC-3-ablated tumours had significantly reduced nuclear staining (Fig. 4e, f). Nuclear-localized pSRC-3-Ser857 represents active SRC-3 in the tumour that in turn promotes target gene expression to maintain tumour growth and metastasis. Importantly, this single phosphorylation site modification was also found to be an indicator of tumour metastasis mediated by ERK3 in a previous study[29]. Taken together, our data demonstrate that

the PFKFB4–SRC-3 signalling axis promotes tumour cell proliferation by increasing purine synthesis (Extended Data Fig. 9g), which may also serve as a critical determinant of metastatic progression of the disease.

To identify the clinical implications of this axis, we first analysed expression of PFKFB4 in The Cancer Genome Atlas (TCGA) database and found its expression to be significantly enhanced across all subtypes of breast cancer (Extended Data Fig. 9h). Because SRC-3 is an ER coactivator, we analysed expression of pSRC-3-Ser857 and PFKFB4 in ER-positive primary breast tumours and adjacent normal tissues. Our data show increased levels of pSRC-3-Ser857, PFKFB4 and SRC-3 in most tumours compared to normal tissues (Extended Data Fig. 10a, b), and a significant correlation between pSRC-3-Ser857 and PFKFB4 levels ($r = 0.63$, Extended Data Fig. 10c). Because PFKFB4 expression is also increased in other breast tumour subtypes, we performed protein array analyses using MDA-MB-231 cell lysates with suppressed expression of SRC-3 or PFKFB4 protein, and compared the significantly altered protein targets to the control non-targeting shRNA. Our study identified a common proteomic signature by intersecting the significant proteins affected by 'both' SRC-3 and PFKFB4 knockdown (Extended Data Fig. 10d). Imposing the restriction of protein-changes in the 'same direction' we evaluated the correlation of the common PFKFB4–SRC-3 proteomic signature with patient survival in a cohort of specimens from patients with breast cancer for which clinical information was available. We identified that the PFKFB4–SRC-3 common proteomic

signature also is associated with a decreased likelihood of survival in a basal-like-subtype triple-negative patient cohort (Extended Data Fig. 10e). These clinical associations are compatible with our in vivo experimental observations substantiating that the PFKFB4–SRC-3 axis is a molecular powerhouse that propels breast tumorigenesis leading it to an aggressive metastatic disease.

Here we have uncovered an interaction between the glycolytic pathway and the oncogenic activation of the transcriptional coactivator SRC-3. The Warburg effect is known to be one of the most dominant sugar metabolic pathways across cancers generating energy and macromolecules to sustain rapid proliferation and tumour growth. We now find that a glycolytic stimulator, the bifunctional enzyme PFKFB4, also can operate as a protein kinase, at least in actively glycolytic tumours. After glucose uptake, PFKFB4 catalyses the synthesis of F2,6BP from F6P and ATP; and our study revealed that under these conditions, PFKFB4 can also phosphorylate SRC-3 at Ser857. Phosphorylation of SRC-3 at Ser857 rapidly increases its transcriptional activity and promotes the synthesis of genes for driving glucose flux towards purine synthesis (Extended Data Fig. 10f). The PFKFB4–SRC-3 axis was found to be enriched in ER-positive breast tumours, and was also identified to promote a common proteomic signature that correlates with worse outcomes in patients with triple-negative breast cancer, thereby driving an aggressive metastatic disease (Extended Data Fig. 10g). Our work suggests that targeting the PFKFB4–SRC-3 axis may be therapeutically valuable in breast tumours that are notably dependent on glucose metabolism.

Online content

1. Ward, P. S. & Thompson, C. B. Metabolic reprogramming: a cancer hallmark even Warburg did not anticipate. *Cancer Cell* **21**, 297–308 (2012).
2. Dang, C. V. Cancer cell metabolism: there is no ROS for the weary. *Cancer Discov.* **2**, 304–307 (2012).
3. Gojis, O. et al. The role of SRC-3 in human breast cancer. *Nat. Rev. Clin. Oncol.* **7**, 83–89 (2010).
4. Anzick, S. L. et al. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* **277**, 965–968 (1997).
5. Xu, J., Wu, R. C. & O'Malley, B. W. Normal and cancer-related functions of the p160 steroid receptor co-activator (SRC) family. *Nat. Rev. Cancer* **9**, 615–630 (2009).
6. Han, S. J., Lonard, D. M. & O'Malley, B. W. Multi-modulation of nuclear receptor coactivators through posttranslational modifications. *Trends Endocrinol. Metab.* **20**, 8–15 (2009).
7. Wu, R. C. et al. Selective phosphorylations of the SRC-3/AIB1 coactivator integrate genomic reponses to multiple cellular signaling pathways. *Mol. Cell* **15**, 937–949 (2004).
8. Wu, R. C., Feng, Q., Lonard, D. M. & O'Malley, B. W. SRC-3 coactivator functional lifetime is regulated by a phospho-dependent ubiquitin time clock. *Cell* **129**, 1125–1140 (2007).
9. Dasgupta, S., Lonard, D. M. & O'Malley, B. W. Nuclear receptor coactivators: master regulators of human health and disease. *Annu. Rev. Med.* **65**, 279–292 (2014).
10. Lonard, D. M. & O'Malley, B. W. The expanding cosmos of nuclear receptor coactivators. *Cell* **125**, 411–414 (2006).
11. Dasgupta, S. & O'Malley, B. W. Transcriptional coregulators: emerging roles of SRC family of coactivators in disease pathology. *J. Mol. Endocrinol.* **53**, R47–R59 (2014).
12. Lonard, D. M., Nawaz, Z., Smith, C. L. & O'Malley, B. W. The 26S proteasome is required for estrogen receptor-α and coactivator turnover and for efficient estrogen receptor-α transactivation. *Mol. Cell* **5**, 939–948 (2000).
13. Yi, P. et al. Atypical protein kinase C regulates dual pathways for degradation of the oncogenic coactivator SRC-3/AIB1. *Mol. Cell* **29**, 465–476 (2008).
14. Pilkis, S. J., Claus, T. H., Kurland, I. J. & Lange, A. J. 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase: a metabolic signaling enzyme. *Annu. Rev. Biochem.* **64**, 799–835 (1995).
15. Rousseau, G. G. & Hue, L. Mammalian 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: a bifunctional enzyme that controls glycolysis. *Prog. Nucleic Acid Res. Mol. Biol.* **45**, 99–127 (1993).
16. Chesney, J. et al. Fructose-2,6-bisphosphate synthesis by 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 4 (PFKFB4) is required for the glycolytic response to hypoxia and tumor growth. *Oncotarget* **5**, 6670–6686 (2014).
17. Chen, H. et al. Nuclear receptor coactivator ACTR is a novel histone acetyltransferase and forms a multimeric activation complex with P/CAF and CBP/p300. *Cell* **90**, 569–580 (1997).
18. Colosia, A. D. et al. Induction of rat liver 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase mRNA by refeeding and insulin. *J. Biol. Chem.* **263**, 18669–18677 (1988).
19. Zhang, C. S. et al. Fructose-1,6-bisphosphate and aldolase mediate glucose sensing by AMPK. *Nature* **548**, 112–116 (2017).
20. Jiang, S. et al. Scaffold attachment factor SAFB1 suppresses estrogen receptor α-mediated transcription in part via interaction with nuclear receptor corepressor. *Mol. Endocrinol.* **20**, 311–320 (2006).
21. Stashi, E. et al. SRC-2 is an essential coactivator for orchestrating metabolism and circadian rhythm. *Cell Reports* **6**, 633–645 (2014).
22. DeBerardinis, R. J., Sayed, N., Ditsworth, D. & Thompson, C. B. Brick by brick: metabolism and tumor cell growth. *Curr. Opin. Genet. Dev.* **18**, 54–61 (2008).
23. Nikolai, B. C. et al. HER2 signaling drives DNA anabolism and proliferation through SRC-3 phosphorylation and E2F1-regulated genes. *Cancer Res.* **76**, 1463–1475 (2016).
24. Patra, K. C. & Hay, N. The pentose phosphate pathway and cancer. *Trends Biochem. Sci.* **39**, 347–354 (2014).
25. Li, L., Zhou, X., Ching, W. K. & Wang, P. Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines. *BMC Bioinformatics* **11**, 501 (2010).
26. Zhu, B. et al. A cell-autonomous mammalian 12 hr clock coordinates metabolic and stress rhythms. *Cell Metab.* **25**, 1305–1319.e9 (2017).
27. Han, J. et al. ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat. Cell Biol.* **15**, 481–490 (2013).
28. Ben-Sahra, I., Hoxhaj, G., Ricoult, S. J. H., Asara, J. M. & Manning, B. D. mTORC1 induces purine synthesis through control of the mitochondrial tetrahydrofolate cycle. *Science* **351**, 728–733 (2016).
29. Long, W. et al. ERK3 signals through SRC-3 coactivator to promote human lung cancer cell invasion. *J. Clin. Invest.* **122**, 1869–1880 (2012).

Author contributions S.D. and B.W.O. conceived the project. S.D. performed most experiments with the following exceptions: N.P. independently performed metabolic profiling and isotope-tracing analysis, J.M.C. and S.Y.J. independently performed proteomics analysis and identified the phosphorylation site, and B.Z. analysed the cistromes and motif analysis. K.R. performed all the data analysis, statistical tests and human datasets for clinical correlation under the supervision of C.C. B.C.N. developed plasmids used in the study; C.E.F. validated the pSRC-3-Ser857 antibody, and P.Y. purified full-length SRC-3 protein and recombinant SRC-3 fragments. T.F.W. provided the RNAi-kinome library and supervised the library screen. X.H.-F.Z. supervised the animal study. B.W.O., M.-J.T. and S.Y.T. supervised the overall project. S.D. and B.W.O. wrote the manuscript with feedback from all authors.

Competing interests The authors declare no competing interests.

METHODS

Vectors and virus production. Commercially-available shRNAs targeting the 3′ UTR region of the *PFKFB4* (TRCN0000199909-sh09 and TRCN0000199820-sh20) and *SRC-3* (TRCN0000370321-sh21 and TRCN0000365196-sh96) were obtained from Sigma. Lentiviruses were produced by transient transfection using Lipofectamine 2000 (Life Technologies) into 293 T cells along with pMD2.G (a gift from D. Trono, Addgene plasmid 12259) and psPAX2 (a gift from D. Trono, Addgene plasmid 12260), and the viral supernatants were collected after 48 h followed by precipitation and purification using PEG-it Virus Concentration Solution (System Bioscience)[30]. The construct expressing the GAL4 responsive luciferase reporter (pG5-luc) was obtained from Promega, and the pBIND-SRC-3 construct was generated by inserting an in-frame fusion between the GAL4 DNA-binding domain and the open reading frame of human *SRC-3*, as previously described[12]. The pBIND-SRC-3(S857A)[31] and pBIND-SRC-3(S857E) mutant were generated using the QuikChange Lightening site-directed mutagenesis kit, as described earlier[31]. The GST–SRC-3 fragment constructs were obtained by cloning portions of the SRC-3 in-frame with GST. The N terminus bHLH (amino acids 1–320), serine/threonine (S/T) (amino acids 321–580), RID (amino acids 581–840), CID (amino acids 841–1080), and HAT (amino acids 1081–1421) domains were generated as previously described[7]. The expression plasmid encoding SRC-3 with a C-terminal Flag tag was cloned into pSG5-Flag (WT SRC-3), and the point mutation of serine 857 to alanine (Ser857Ala) was generated by site-directed mutagenesis of wild-type SRC-3 and GST–SRC-3-CID constructs[13,29]. The human *PFKFB4* cDNA (NM_004567.3) was obtained from Origene (RC201573). The PFKFB4 mutants Gly46Ala, Pro48Ala, Gly51Ala, Arg230Ala and Arg238Ala were generated by site-directed mutagenesis. All constructs were verified by Sanger sequencing. The siGENOME siRNA against *PFKFB4*, *SRC-3* and *ATF4* were obtained from Dharmacon.

The shRNA sequences were as follows: shPFK#09 (TRCN0000199909): 5′-CCGGGCTGATTGGCTGCCACATTTCCTCGAGGAAATGTGGCAGCCAATCAGCTTTTTTG-3′; shPFK#20 (TRCN0000199820): 5′-CCGGGCGCAGCTCTTAGGTGTTCACCTCGAGGTGAACACCTAAGAGCTGCGCTTTTTTG-3′; shSRC-3#21 (TRCN0000370321): 5′-CCGGTGACACTGCACTAGGATTATTCTCGAGAATAATCCTAGTGCAGTGTCATTTTTG-3′; shSRC-3#96 (TRCN0000365196): 5′- CCGGTTCCACCTCCTAGGGATATAACTCGAGTTATATCCCTAGGAGGTGGAATTTTTG-3′.

Cell culture. HeLa, HEK293T, MDA-MB-231, MCF-7 and MCF-7-ERE-MAR-Luc cells were cultured in DMEM (Gibco) supplemented with 10% FBS; SK-BR-3 cells were grown in McCoy's medium with 10% FBS; and MCF-10A cells in DMEM/F12 (Gibco) supplemented with 5% horse serum, epidermal growth factor (EGF), hydrocortisone, cholera toxin and insulin. All cell lines were incubated at 37 °C and 5% CO$_2$. Cell lines were obtained from ATCC, and maintained and yearly tested for mycoplasma contamination by the Tissue Culture Core, Baylor College of Medicine (BCM).

Stable cells expressing shRNAs were generated by lentiviral transduction in presence of polybrene (8 μg ml^{-1}). Polyclonal pooled populations of stable cells were selected in the presence of puromycin (1 μg ml^{-1}) for more than three passages before initiating any functional experiments.

Human kinome library screen. A high-throughput RNAi screen was performed using the Stealth RNAi human kinase library (Life Technologies) targeting each 636 human kinases with three individual siRNAs directed at different regions of the gene that were arrayed in twenty-four 96-well plates. To identify the kinases that modulate SRC-3 transcriptional activity, we reverse co-transfected HeLa cells with pBIND or pBIND-SRC-3 (2 ng per 96-well) along with pG5luc firefly-luciferase reporter (100 ng per 96-well) and control siRNA targeting GFP (siGFP) or siRNAs targeting kinases (40 nM). The mixture was incubated with 0.75 μl per 96-well of Lipofectamine 2000 for 20 min followed by addition of HeLa cell suspension (12,500 cells per 96-well) in complete growth medium (DMEM plus 10% FBS) on top. After 48 h of culture, plates were carefully washed with PBS and luminescence reading was recorded in luminometer (Berthold) using the Dual-Luciferase Assay System (Promega). Additional wells present on all plates had appropriate controls containing cells transfected with pBIND and siRNA targeting GFP (siGFP), or pBIND-SRC-3 and siGFP along with reporter plasmid. SRC-3 transcriptional activity was calculated by comparing the relative luciferase units (RLU) of pBIND-SRC-3 to pBIND readings transfected with siGFP. Firefly luciferase reading from each well was normalized to its *Renilla* reading (pBIND-vector backbone contains the *Renilla* luciferase gene) to adjust the variations in transfection efficiency. The fold change in SRC-3 activity upon suppression of kinases was calculated by comparing data to siGFP readings, followed by robust *z*-score analysis to identify kinases that either increase or decrease SRC-3 activity more than 2 s.d. above or below control siGFP (pBIND-SRC-3 plus siGFP) treatment. Fold change values were converted to log$_2$ for each set of siRNA and then graphed in 3D plot.

Cell proliferation assays. Cells were transfected with indicated siRNAs and were seeded at a density of 3,000 cells per 96-well in complete growth medium.

For rescue experiment, cells were seeded in complete growth medium supplemented with dialysed serum with or without purines (10 μM adenosine, Sigma and 10 μM guanosine, Sigma). After 4 days, cells were stained with CellTiter96 (Promega) reagent followed by measurement of absorbance at 490 nm. For the clonogenic survival assays, 1,000 cells per well were plated onto a 6-well plate, and were incubated for 7 days, and stained with crystal violet. The medium was changed every two days.

In vitro phosphorylation assays. The full length SRC-3–Flag protein was expressed in Sf9 cells and purified using anti-Flag antibody beads[13]. The SRC-3 fragments were expressed as GST fusion proteins in *Escherichia coli* and purified using a GST fusion protein purification kit (Life Technologies) following the manufacturer's protocol. Each reaction of the in vitro phosphorylation assay was carried out with varying concentration of purified recombinant GST–PFKFB4 protein (SignalChem) (0.1–1 μg) along with SRC-3 (0.25 μg) as substrate, cold ATP (0.2 mM) or 5 μCi [γ-^{32}P]ATP (Perkin Elmer), and 1 × kinase buffer (Cell Signaling) in a total volume of 30 μl. The reaction was carried out at 30 °C for 30 min and then stopped by adding 10 μl of 4× SDS sample buffer. Proteins were resolved by SDS–PAGE gel, stained with Coomassie blue (Bio-Rad), and visualized by autoradiography or probed with anti-Ser/Thr antibody (BD Biosciences). For mass spectrometric (MS) identification of phosphorylation sites, the GST–SRC-3-CID protein was used as a substrate for the kinase reaction along with cold ATP and PFKFB4 enzyme followed by separation by SDS–PAGE and staining with Coomassie blue. Gel lanes were sliced into different bands and in-gel digested overnight at 37 °C with trypsin. After digestion, peptides were extracted twice in 200 μl of acetonitrile with re-suspension in 20 μl of 2% formic acid before second extraction, dried in a Savant SpeedVac, and dissolved in a 5% methanol, 0.1% formic acid solution. The samples were then subjected to mass spectrometry separation to detect phosphorylated residues.

Cell culture treatment conditions, protein isolation and immunoblotting. For siRNA treatments, cells were lysed 72 h after transfection. Stable cells were grown until 80% confluency before protein was extracted. During nutritional stress conditions, stable cells were cultured in complete medium until 80% confluency, followed by a brief starvation (3 h) in glucose-free growth medium. Cells were then switched to glucose-free DMEM supplemented with 10% dialysed serum and 5 mM or 25 mM glucose, as indicated in the figures, for 24 h before cells were lysed. For glucose withdrawal, cells were cultured in 25 mM for 24 h and then switched to medium containing 5 mM glucose for an additional 6 h. For FBP treatment, glucose-starved cells were pre-treated with 10 μM streptolysin O (Sigma, S5265) to permeabilize the cells, followed by the addition of FBP (Santa Cruz, sc-214805) as previously described[19]. Immunoblotting was performed as previously described[30]. In brief, cells were lysed using NP-40 lysis buffer (Life Technologies) along with protease and phosphatase inhibitor cocktail (Millipore). Total protein was estimated using a BCA protein estimation kit (Pierce) and approximately 40 μg of proteins were separated by 4–12% Bis-Tris gels (Life Technology) and electroblotted onto nitrocellulose membranes using the iBlot system (Life Technology). Blots were blocked for 2 h at room temperature or overnight at 4 °C in 1 × TBS buffer (Biorad) supplemented with 0.1% Tween-20 (Sigma) and either 5% bovine serum albumin (BSA) or 5% non-fat dry milk (Biorad). Blots were incubated overnight at 4 °C with primary antibody diluted into TBST containing 1% BSA or 5% non-fat dry milk. Blots were subsequently washed three times for 10 min in TBST and incubated with secondary antibody coupled to HRP (Promega). Blots were washed as previously described, reacted with ECL reagents (Thermo Fisher Scientific) and detected by chemi-luminescence (UVP Biospectrum). Semi-quantitative levels of each band were analysed by densitometry using UVP Vision Works LS software, and the relative values normalized to actin are indicated numerically under each lane.

Antibodies used for immunoblotting in the study are: mouse monoclonal SRC-3 (611105, BD Biosciences), rabbit monoclonal SRC-3 (2126, Cell Signaling), Flag (F3165, Sigma-Aldrich), mouse phosphoSerine/Threonine (612548, BD Biosciences), rabbit PFKFB4 (137785 and 71622, Abcam), mouse PFKFB4 (TA500809, Origene), rabbit monoclonal ATF4 (11815, Cell Signaling), and β-actin conjugated to HRP (A3854, Sigma-Aldrich). The phospho-SRC-3 (Ser857) rabbit monoclonal antibody was cell culture supernatant produced from hybridoma generated by immunizing animals with a synthetic peptide containing phosphorylated Ser857 of human SRC-3. This antibody (Clone 10A6) was a gift from Cell Signaling Technology.

Immunoprecipitations. 293 T cultured in 100 mm dishes until 80% confluency was transfected with Flag–SRC-3 followed by infection with adenovirus PFKFB4 (Signagen Laboratories). Twenty-four hours after infection, the medium was changed and cells were incubated overnight in different concentrations of glucose (5, 10, 15 and 25 mM) in glucose-free DMEM medium supplemented with 10% dialysed FBS. For MDA-MB-231 cells, stable cells expressing shRNAs targeting *PFKFB4* or *SRC-3* or expressing Ser857Ala in SRC-3-depleted cells were grown in 5 mM or 25 mM glucose. Cells were lysed in NP-40 lysis buffer (Invitrogen)

supplemented with protease and phosphatase inhibitor cocktail (Millipore). For co-immunoprecipitations, lysates were precleared with control Protein A/G Agarose beads (Pierce). Five hundred micrograms of protein were then used for pull-down assays using monoclonal anti-Flag (F3165, Sigma) or anti-phosho-Ser/Thr antibody (BD Biosciences) overnight. The beads were then captured, washed and immunoprecipitated proteins were eluted and subjected to immunoblotting, along with 2% input sample run in parallel. For ATF4 pull-down, anti-ATF4 (11815, Cell Signaling) was used at a 1:250 dilutions to pull down ATF4. Light-chain-specific anti-rabbit secondary antibody conjugated to HRP (Jackson Immunoresearch, 1:5,000) was used to detect ATF4 in immunoblotting following immunoprecipitation to avoid overlap with IgG-heavy chain

Immunohistochemistry. Immunohistochemistry was performed as previously described[30]. Mouse monoclonal anti-human Ki-67 antibody MIB-1 (Dako) and rabbit monoclonal anti-phospho-SRC-3 (Ser857) (Cell Signaling) were used to stain the lung sections followed by anti-mouse or anti-rabbit Alexa-594 secondary antibody (Molecular Probes).

Gene expression analyses. Total RNA was isolated from cancer cells or tumours using the RNeasy Kit (QIAGEN). Reverse transcription was carried out using a Superscript VILO cDNA synthesis kit (Invitrogen) according to the manufacturer's instructions. For gene expression analysis, qPCR was performed using the Taqman system (Roche) with sequence-specific primers and the Universal Probe Library (Roche). *ACTB* was used as an internal control. Melt curve analysis was performed to ensure that a single PCR product was produced in a given well. We used three biological replicates for each treatment group. Data were analysed using the comparative C_t method ($\Delta\Delta C_t$).

Targeted TCA, glycolysis, PPP and nucleotide synthesis, and intermediary metabolite analysis using liquid chromatography–mass spectrometry. Sample preparation for mass spectrometric analysis: the metabolome extraction method described earlier was used for the cell lines in this study[30,32]. In brief, cells were thawed at 4 °C and subjected to freeze–thaw cycles in liquid nitrogen three times to rupture the cell membrane. Following this, 750 μl of ice-cold methanol:water (4:1) containing 20 μl of spiked internal standard was added to each cell line. The cells were homogenized for 1 min (30 s pulse twice) and mixed with 450 μl of ice-cold chloroform and vortex mixed in a Multi-Tube Vortexer for 10 min. The resulting homogenate was mixed with 150 μl of ice-cold water and vortexed again for 2 min. The homogenate was incubated at −20 °C for 20 min and centrifuged at 4 °C for 10 min to partition the aqueous and organic layers. The aqueous and organic layers were separated and dried at 37 °C for 45 min in an Automatic Environmental Speed Vac system (Thermo Fisher Scientific). The aqueous extract was reconstituted in 500 μl of ice-cold methanol:water (50:50) and filtered through 3 kDa molecular filter (Amicon Ultracel-3K Membrane, Millipore Corporation) at 4 °C for 90 min to remove proteins. The filtrate was dried at 37 °C for 45 min in a speed vac and stored at −80 °C until mass spectrometry analysis. Before mass spectrometry analysis, the dried extract was re-suspended in 100 μl of methanol:water (50:50) containing 0.1% formic acid and analysed using multiple reaction monitoring (MRM).

Liquid chromatography–mass spectrometry HPLC analysis was performed using an Agilent 1290 series HPLC system equipped with a degasser, binary pump, thermostatted auto sampler and column oven (all from Agilent Technologies). The MRM-based measurement of relative metabolite levels were used for normal phase chromatographic separation. All samples were kept at 4 °C, and 5 μl of the sample was used for analysis.

Separation of TCA, glycolysis and PPP-associated metabolites. The normal phase chromatographic separation was also used for targeted identification of metabolites. This analysis used solvents containing water (solvent A), with solvent A modified by the addition of 5 mM ammonium acetate (pH 9.9), and 100% acetonitrile (solvent B). The binary pump flow rate was 0.2 ml min⁻¹ with a gradient spanning 80% B to 2% B over a 20-min period followed by 2% B to 80% B for a 5-min period and followed by 80% B for a 13-min time period. The flow rate was gradually increased during the separation from 0.2 ml min⁻¹ (0–20 min), 0.3 ml min⁻¹ (20.1–25 min), 0.35 ml min⁻¹ (25–30 min), 0.4 ml min⁻¹ (30–37.99 min) and finally set at 0.2 ml min⁻¹ (5 min). Metabolites were separated on a Luna Amino (NH2) column (4 μm, 100 A 2.1 × 150 mm, Phenominex) that was maintained in a temperature-controlled chamber (37 °C). All the columns used in this study were washed and reconditioned after every 50 injections. Ten microlitres was injected and analysed using a 6495 QQQ triple quadrupole mass spectrometer (Agilent Technologies) coupled to a 1290 series HPLC system via selected reaction monitoring (SRM). Metabolites were measured using negative ionization mode with an electrospray ionization (ESI) voltage of −3,500 eV, respectively. Approximately 9–12 data points were acquired per detected metabolite.

Separation of nucleotides. For measurement of nucleotides and deoxy-nucleotides before mass spectrometry analysis, the dried extract was suspended in 50 μl of methanol:water (50:50) containing 0.1% formic acid. Samples were delivered to the MS via reverse phase chromatography using a RRHD SB-CN column (1.8 μm, 3.0 × 100 mm, Agilent Technologies) at 300 μl min⁻¹. The gradient spanned

2% B to 98% B over a 15-min period followed by 98% B to 2% B for a 1-min period. The gradient was continued for a 4-min time period to re-equilibrate the column. Buffers A and B consisted of 0.1% formic acid in water and acetonitrile, respectively.

Ten microlitres was injected and analysed using a 6495 QQQ triple quadrupole mass spectrometer (Agilent Technologies) coupled to a 1290 series HPLC system via SRM. Metabolites were measured using positive ionization mode with an ESI voltage of 4000 eV, respectively. Approximately 9–12 data points were acquired per detected metabolite.

Isotope labelling and profiling by targeted mass spectrometry. Glucose labelled with [6-¹³C]glucose and [U-¹³C]glucose were purchased from Cambridge Isotope Laboratories. MDA-MB231 cells were grown in 10-cm dishes in regular medium until 80% confluence, followed by brief (3 h) starvation and then addition of 25 mM of [6-¹³C]glucose supplemented with glucose-free DMEM medium with 10% dialysed FBS and 1% penicillin/streptomycin[33]. For [U-¹³C]glucose, cells were fed with steady-state isotope tracers for 48 h and medium was replaced 2 h before metabolome collection and/or isotope tracer addition. Culture medium was collected, cells were washed with PBS, counted, and snap-frozen in liquid nitrogen. Cells were scraped into a 0.5-ml mixture of 1:1 water:methanol, sonicated for 1 min (two 30-s pulses), and then mixed with 450 μl ice-cold chloroform. The resulting homogenate was then mixed with ice-cold water and vortexed again for 2 min. The homogenate was incubated at −20 °C and centrifuged at 4 °C for 10 min to partition the aqueous and organic layers. The aqueous and organic layers were combined and dried at 37 °C for 45 min in an automatic Environmental Speed Vac system (Thermo Fisher Scientific). The extract was reconstituted in a 500 μl solution of ice-cold methanol:water (1:1) and filtered through a 3-kDa molecular filter (Amicon Ultracel 3-kDa Membrane) at 4 °C for 90 min to remove proteins. The filtrate was dried at 37 °C for 45 min in a speed vacuum and stored at −80 °C until mass spectrometry analysis. Before MS analysis, the dried extract was resuspended in a 50 μl solution of methanol:water (1:1) containing 0.1% formic acid, and then analysed using MRM. Ten microlitres was injected and analysed using a 6490 QQQ triple quadrupole mass spectrometer (Agilent Technologies) coupled to a 1290 Series HPLC system via SRM. Metabolites were targeted in both positive and negative ion modes: the ESI voltage was 4,000 V in positive ion mode and –3,500 V in negative ion mode. Approximately 9–12 data points were acquired per detected metabolite. To target the TCA flux, the samples were delivered to the mass spectrometer via normal-phase chromatography using a Luna Amino column (4 μm, 100 A 2.1 × 150 mm). To target the fatty-acid flux, the samples were delivered to the mass spectrometer via reverse-phase chromatography using a Phenyl Hexyl column (3 μm, 100 A 2.1 × 150 mm). For ¹³C-labelled experiments, SRM was performed for expected ¹³C incorporation in various forms for targeted liquid chromatography–tandem mass spectrometry (LC–MS/MS). Mass isotopomer distribution (MID) was calculated and corrected for natural abundance.

Proximity ligation assay. Interaction between endogenous SRC-3 and PFKFB4 was detected using the PLA technique[34] using Duolink In situ Red Starter Kit Mouse/Rabbit (UO92101, Sigma) according to manufacturer's instructions. In brief, MDA-MB-231 cells were seeded in a 35-mm glass bottom culture dish (P35G-0-14C, MatTek Corporation), and after reaching 80% confluency, cells were fixed followed by blocking for 1 h using the Duolink Blocking Solution at 37 °C. Cells were then incubated in presence of primary antibodies: SRC-3 (rabbit monoclonal, Cell Signaling) and PFKFB4 (mouse monoclonal, Origene), either alone or in combination. After incubation, cells were washed and Duolink PLA PLUS and MINUS probes were added for 1 h at 37 °C. After washing off the unbound probes, cells were incubated first with the ligase enzyme followed by DNA polymerase enzyme to amplify the DNA circle. Finally, cells were mounted using Duolink In Situ Mounting Media with DAPI, and analysed by microscopy. Images were obtained using Zeiss Axio Observer A1 inverted microscope with N-Achroplan 100 × /1.25 oil lens, Zeiss MRC5 camera, and AxioVision Rel.4.8 software.

Analysis of ATF4 and SRC-3 cistromes and motif analysis of ATF4-bound sequences. Owing to the lack of ATF4 and SRC-3 ChIP–seq datasets in breast cancer cell lines, we compared an in-house SRC-3 ChIP–seq dataset of mouse liver[26], with previously published ATF4 ChIP–seq data in mouse embryonic fibroblasts[27]. Even though this comparison is less than ideal as SRC-3 and ATF4 ChIP–seq were performed in different tissues, the co-localization of SRC-3 and ATF4 cistromes even in different tissues, nevertheless, argues for an interplay between them, a finding subsequently confirmed by co-immunoprecipitation and ChIP–qPCR assay in human breast cancer cell lines. ATF4 binding motifs in the promoter regions of *XDH*, *TKT* and *AMPD1* genes were discovered using the MISP (Motif-based Interval Screener with PSSM) toolbox in Galaxy Cistrome with a *P* value cut-off of 0.005. The consensus ATF4-binding motif used as input is TGATGCAA.

ChIP. The following antibodies were used for ChIP: SRC-3 (Cell Signaling or BD Biosciences), ATF4 (Santa Cruz C-20, and 11815 Cell Signaling), pSRC-3-S857 (Cell Signaling), and rabbit IgG. ChIP assays were performed according to an EZ ChIP kit (Millipore) with some modification[35]. In brief, MDA-MB-231 cells were

grown in 15-cm dishes until 80% confluent. For glucose stimulation, cells were glucose-deprived for 3 h by incubating in glucose-free DMEM supplemented with 10% FBS, followed by 4 h stimulation with 5 mM or 25 mM glucose. Cells were cross-linked in 1% formaldehyde and quenched with 125 mM glycine. Chromatin was sheared by sonication using a Branson Sonifier, precleared with control IgG antibodies and agarose beads (Millipore), and then immunoprecipitated with IgG (control), SRC-3, pSRC-3-S857 and ATF4 antibodies. DNA fragments were eluted from beads followed by reverse-crosslinking and purified DNA was used in qPCR reactions using SYBR green (Applied Biosystems) to determine the promoter occupancy. Melt curve analysis was performed to verify all SYBR green reactions produced a single PCR product.

Luciferase assays. Luciferase assays were performed from whole-cell lysates made in Cell Culture Lysis reagent (Promega) using the Luciferase Reporter Assay (Promega) and a Berthold 96-well plate reader. Luciferase values were normalized to the total protein level.

Metabolomic phenotyping microarrays. Screening was performed using 96-well plate phenotype microarrays (Biolog) containing 88 different carbon substrates and 5 nucleotides as the energy source[21]. MCF10A cells were infected with adenovirus expressing GFP or SRC-3, and seeded at an initial density of 2×10^4 cells per well in triplicate. Biolog Redox Dye Mix MA was added to each well according to the manufacturer's instructions, and kinetic usage of the metabolites was monitored using the GEN III OmniLog ID System (Biolog).

Human breast tumours. The breast tumours and adjoining normal tissue was obtained from the Lester and Sue Smith Breast Center at Baylor College of Medicine according to the Institutional Research Board approved protocol #H-7900. Whole-cell lysates from a total of 14 human breast tumours that are ER+ primary tumours, along with matched normal tissues, were used to detect pSRC-3-Ser857, SRC-3, and PFKFB4 levels by immunoblotting.

Determining a common PFKFB4–SRC-3 proteomic signature. Protein lysates from MDA-MB-231 cells stably expressing shRNAs targeting *SRC-3* or *PFKFB4* were used for protein array analysis as described before[36]. Expression of proteins significantly altered owing to the ablation of PFKFB4 and SRC-3 compared with non-targeting control shRNA were determined using a parametric *t*-test as implemented in the python (spicy) statistical system. Significance was assessed for $P < 0.05$, fold change exceeding $1.25 \times$, and normalized signal levels exceeding 200 U. A common proteomic signature was determined by intersecting the significant proteins affected by each treatment, and imposing the restriction that the protein changes are in the same direction.

Association of the PFKFB4–SRC-3 proteomic signature in human basal breast cancer. We evaluated the association of the common PFKFB4–SRC-3 proteomic signature with patient survival in a cohort of primary basal breast cancer patient specimens collected by The Cancer Genome Atlas (TCGA) for which clinical information has been collected[37]. We first subsetted the proteins measured using the array by TCGA. Next, for each protein in the PFKFB4–SRC-3 common proteomic signature and for each basal breast cancer specimen, we computed the *z*-score for its expression within the patient cohort. We then computed the sum of the *z*-scores for each specimen. Specifically, the *z*-scores of the proteins suppressed by PFKFB4–SRC-3 (that is, upregulated by *PFKFB4* and *SRC-3* shRNA) were subtracted from the *z*-scores of the proteins induced by PFKFB4 (that is, downregulated by *PFKFB4* and *SRC-3* shRNA); this resulted in an activity score of the PFKFB4–SRC-3 common proteomic signature, respectively, for each specimen. After computing the activity scores, we further partitioned the patient cohort into specimens with a high activity score (top 33% of the specimens) and specimens with a low activity score (bottom 33% of the specimens) for the corresponding signatures.

We considered significant association with survival using the log-rank test ($P < 0.05$) and the Cox proportional hazard test ($P < 0.05$) available via the package survival as implemented in the R statistical system.

Tumorigenicity and metastasis assays. All animal experiments were carried out in accordance with a protocol approved by the Baylor College of Medicine Institutional Animal Care and Use Committee and experiments were terminated once maximal tumour volumes were reached (10% of the animal body weight). MDA-MB-231 breast cancer cells stably expressing luciferase were individually transduced with shRNAs targeting *SRC-3* and *PFKFB4*. For the rescue experiment, SRC-3-ablated tumour cells were used to restore the levels of either wild-type SRC-3 or the SRC-3(Ser857Ala) mutant, and the polyclonal pooled population was selected. Approximately 2.5×10^5 cells were injected at orthotopic site along with Matrigel (BD Biosciences) (1:1 volume) in the mammary fat pad of 5–6 week-old female athymic nude Foxn1-nu mice (Envigo). The mammary tumour length (L) and width (W) were measured with a caliper. Tumour volumes were calculated using the formula $L \times W^2 \times \pi/6$. After six weeks, tumours were resected out by surgery and the animals were monitored for lung metastasis progression every week and quantified using noninvasive bioluminescence measurement with IVIS Lumina II equipment. Four weeks after tumour resection animals were euthanized and tissues were collected and fixed in 4% PFA. Paraffin-embedded lung samples were also subjected to haematoxylin and eosin staining to reveal the size and number of lung macro or micro-metastases. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. No statistical methods were used to predetermine sample size estimate.

Statistics. Unless otherwise indicated, all results represent the mean ± s.d., and statistical comparisons between different groups were performed using the two-tailed Student's *t*-test, one-way or two-way ANOVA with appropriate multiple comparisons corrections. For all statistical analyses, differences of $P \leq 0.05$ were considered statistically significant, and three biologically independent experiments with similar results are reported. GraphPad Prism software version 6.0/7.0 (GraphPad Software) was used for data analysis.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The ChIP–seq data have been submitted to the Gene Expression Omnibus under accessions GSE35681 (for ATF4) and is GSE67860 (for SRC-3). Other data that support the findings of this study are available from the corresponding author upon reasonable request.

30. Dasgupta, S. et al. Coactivator SRC-2-dependent metabolic reprogramming mediates prostate cancer survival and metastasis. *J. Clin. Invest.* **125**, 1174–1188 (2015).
31. Wang, L. et al. Characterization of a steroid receptor coactivator small molecule stimulator that overstimulates cancer cells and leads to cell stress and death. *Cancer Cell* **28**, 240–252 (2015).
32. Jin, F. et al. A novel [^{15}N]glutamine flux using LC-MS/MS-SRM for determination of nucleosides and nucleobases. *J. Anal. Bioanal. Tech.* **6**, 267 (2015).
33. Fan, J. et al. Quantitative flux analysis reveals folate-dependent NADPH production. *Nature* **510**, 298–302 (2014).
34. Stadler, C. et al. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods* **10**, 315–323 (2013).
35. Foulds, C. E. et al. Proteomic analysis of coregulators bound to ERα on DNA and nucleosomes reveals coregulator dynamics. *Mol. Cell* **51**, 185–199 (2013).
36. Welte, T. et al. Oncogenic mTOR signalling recruits myeloid-derived suppressor cells to promote tumour initiation. *Nat. Cell Biol.* **18**, 632–644 (2016).
37. Ciriello, G. et al. Comprehensive Molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).

a

b

c

d

e

f Secondary Screen

g

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Kinome-wide screen identified potential kinases regulating SRC-3 intrinsic transcriptional activity. a, HeLa cells expressing varying concentrations of the pBIND or pBIND-SRC-3 constructs were used to measure SRC-3 activity. $n = 4$ biologically independent samples. *$P < 0.000001$, one-way ANOVA with Sidak's multiple comparison test. RLU are normalized by protein content. **b**, HeLa cells expressing pBIND or pBIND-SRC-3 were treated with siRNA targeting GFP or *PRKCZ* at the indicated dose followed by luciferase assay to measure SRC-3 activity. $n = 3$ biologically independent samples. *$P < 0.000001$, one-way ANOVA with Tukey's multiple comparison test. **c**, Different control siRNAs targeting GFP or luciferase (Luc) were used to measure SRC-3 activity in HeLa cells expressing pBIND or pBIND-SRC-3. $n = 3$ biologically independent samples. The GFP control siRNAs in the red box were used in the library screen as controls. **d**, Effect on SRC-3 transcriptional activity by three sets of siRNA (sets A, B and C) targeting

636 human kinases in HeLa cells. Effect of GFP control siRNA was set at 1 (dotted line), the cut-off fold for increased activation was set at 2, and reduced activity at 0.75 following *z*-score analysis. $n = 3$ siRNAs per / kinase, $n = 6$ siGFP per plate; total $n = 1,908$ (siRNAs targeting kinases) $n = 144$ (siGFP control) independent samples. **e**, SRC-3 activity in HeLa cells across 24 kinome-library plates in the presence of control siRNA targeting GFP. $n = 6$ biologically independent replicates for each plate. **f**, A secondary screen was performed in HeLa cells to confirm the primary screen hits using a pooled siRNA targeting the kinases followed by SRC-3 transcriptional activity. $n = 3$ biologically independent samples. Boxes are as in Fig. 2e. **g**, Relative proliferation of MDA-MB-231 cells 4 days after treatment with siRNAs targeting GFP (control), *SRC-3* or the indicated kinases. $n = 3$ biological replicates. *$P < 0.0001$. two-way ANOVA with Dunnett's multiple comparisons test. Unless stated otherwise, data are mean ± s.d.
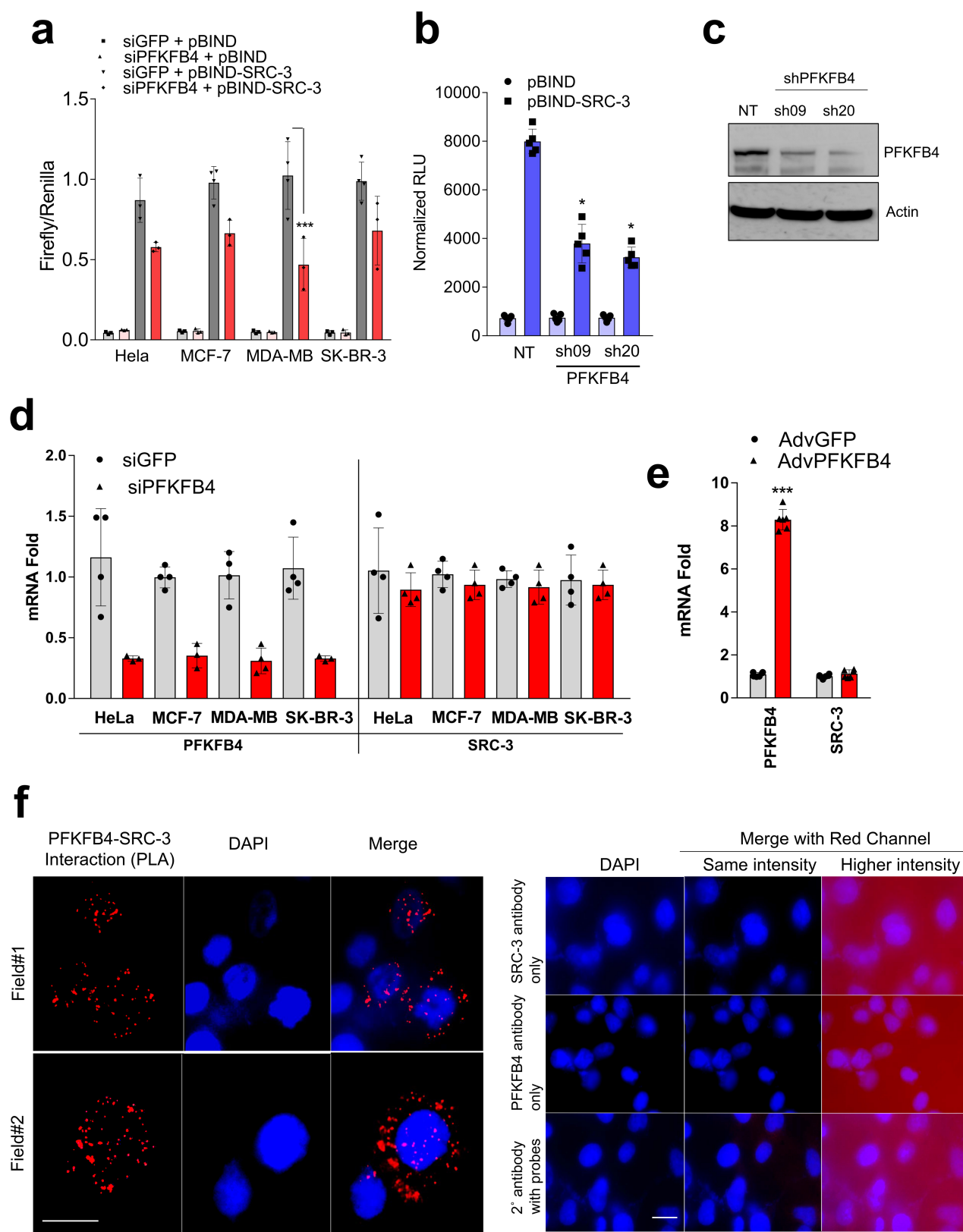
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | PFKFB4, the top hit from the kinase screen, enhances the transcriptional activity of SRC-3. a, Effect of PFKFB4 knockdown on SRC-3 transcriptional activity in various breast cancer cell lines. $n = 3$ or $n = 4$ (siGFP plus pBIND-SRC-3) biologically independent cells. *$P < 0.000009$, two-way ANOVA with Tukey's multiple comparison test. **b**, SRC-3 transcriptional activity in MDA-MB-231 cells expressing shRNAs targeting *PFKFB4* (#09 and #20) or non-targeting control, co-transfected with pBIND or pBIND-SRC-3. $n = 5$, biological replicates. *$P < 0.0001$, one-way ANOVA with Tukey's multiple comparisons test. **c**, Protein expression of PFKFB4 or actin in MDA-MB-231 cells expressing shRNAs targeting *PFKFB4*. **d**, Expression of *PFKFB4* and *SRC-3* mRNA in indicated breast tumour cells after treatment with siRNAs targeting GFP control or *PFKFB4*. $n = 4$ or $n = 3$ biological replicates. See Source Data for exact P values. **e**, Expression of *PFKFB4* and *SRC-3* mRNA in

MDA-MB-231 cells transduced with adenoviruses expressing GFP or PFKFB4. $n = 6$ biologically independent cells. ***$P < 0.000001$, two-way ANOVA with Tukey's multiple comparison test. **f**, Left, MDA-MB-231 cells were stained with specific antibodies against SRC-3 (rabbit) and PFKFB4 (mouse) before proximity ligation assay (PLA). The PLA signals between endogenous SRC-3 and PFKFB4 are shown in the red channel, DAPI was used to stain the nuclei (blue) and the merge images show the overlay of the red and blue channels. Two representative fields from biologically independent experiments were shown from $n = 5$. Right, control cells were stained with either one of the antibodies against SRC-3, PFKFB4 or secondary antibody-conjugated with probes. Scale bars, 20 μm (left), 40 μm (right). Data are representative of three biologically independent experiments with similar results, and are shown as mean ± s.d.

a

| SRC-3 (0.25ug) | - | + | + | + | + | + |
| PFKFB4 (ug) | - | - | 0.1 | 0.25 | 0.5 | 1.0 |

—160kD

IB: p-Ser/Thr

b

| F6P (mM) | 0.25 | 0.5 | 1.0 | - | - | - | 0.25 | 0.5 | 1.0 |
| ATP (mM) | - | - | - | 0.1 | 0.2 | 0.4 | 0.1 | 0.2 | 0.4 |

p-Ser/Thr

PFKFB4

c

SRC-3 FL, bHLH, S/T, RID, CID, HAT

PFKFB4→

d Phosphorylated S857

A V pS L D S P V S V G S S P P V K

[M+2H]²-P

Intensity [counts] [10^3]

m/z

e S857A mutant

A V A L D S P V S V G S S P P V K

Intensity [counts] (10^3)

m/z

f

● NT
■ PFKFB4-Sh#09
▲ PFKFB4-Sh#20

mRNA Fold

PFKFB1 PFKFB2 PFKFB3 PFKFB4

g

NTshRNA, shSRC-3, shSRC-3 + S857A, shSRC-3 + WT-SRC-3

pSRC-3 (Ser-857)
1.0   0.1   0.15   0.7

Actin

SRC-3
1.0   0.1   0.85   0.8

Actin

h

	NT-shRNA			shPFKFB4		
Glucose (25mM)	+	-	-	+	-	-
FBP (10µM)	-	-	+	-	-	+

pSRC-3-S857
1.0   0.4   1.1   0.1   0.15   0.2

SRC-3

Actin

i

RLU

● pBIND+Adv.GFP(-)
▲ pBIND-SRC-3+Adv.GFP(-)
▼ pBIND+Adv.GFP(+)
◆ pBIND-SRC-3+Adv.GFP(+)
● pBIND+Adv.PFKFB4(-)
▲ pBIND-SRC-3+Adv.PFKFB4(-)
▼ pBIND+Adv.PFKFB4(+)
◆ pBIND-SRC-3+Adv.PFKFB4(+)

Adv. GFP / Adv. PFKFB4

****

| Glucose | - | + | - | + | - | + | - | + | - | + | - | + |
| | 5mM | | | | 15mM | | | | 25mM | | | |

Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | PFKFB4 functions as a protein kinase by phosphorylating SRC-3 at the Ser857 residue. a, In vitro PFKFB4 kinase assay in the presence of purified SRC-3 protein, F6P, ATP and increasing concentration of recombinant PFKFB4 enzyme followed by SDS–PAGE. Immunoblotting with pSer/Thr antibody shows the level of phosphorylated SRC-3 protein. **b**, In vitro PFKFB4 kinase assay in presence of purified SRC-3 protein, PFKFB4 enzyme and varying concentrations of F6P and ATP followed by SDS–PAGE. Immunoblotting with pSer/Thr antibody shows the level of pSRC-3 protein. **c**, Coomassie blue stain showing the levels of GST-fused SRC-3 fragments used in in vitro kinase reactions performed in Fig. 2b. **d**, Proteomics analysis of in vitro kinase assay using the GST–SRC-3-CID fragment in the presence of PFKFB4 enzyme and ATP followed by mass spectrometric analyses. Mass spectrum shows the green phosphorylation peak. **e**, Proteomics analysis of an in vitro kinase assay using a Ser857Ala-mutated GST–SRC-3-CID protein in the presence of PFKFB4 enzyme and ATP, followed by mass spectrometric analyses. Mass spectrum failed to detect phosphorylation peaks in the Ser857Ala-mutated SRC-3-CID protein. **f**, Expression of PFKFB1, PFKFB2, PFKFB3 and PFKFB4 in MDA-MB-231 cells expressing shRNAs targeting *PFKFB4* (#09 and #20). mRNA levels were normalized to internal housekeeping gene *ACTB*. $n = 3$ biological replicates. *$P < 0.05$, two-way ANOVA with Tukey's multiple comparisons test. **g**, Protein levels of pSRC-3-Ser857, total-SRC-3 and actin in MDA-MB-231 cells stably expressing non-targeting control shRNA, *SRC-3* shRNA, or *SRC-3* shRNA plus the shRNA-resistant Ser857Ala SRC-3 mutant (shSRC-3 + S857A) or *SRC-3* shRNA plus wild-type SRC-3 (shSRC-3 + WT-SRC-3) cultured in 25 mM glucose. Protein bands were quantified by ImageJ after normalization to β-actin. **h**, MDA-MB-231 cells stably expressing non-targeting shRNA or shRNA targeting *PFKFB4* were grown in the presence of 25 mM glucose or were glucose-starved for 4 h followed by incubation with streptolysin O for 5 min. FBP (10 μM) was added to glucose-starved cells for an additional 1 h, followed by cell lysis and immunoblotting. Protein bands were quantified by ImageJ after normalization to β-actin and the non-targeting shRNA lane was set to 1. **i**, Relative luciferase activity showing the transcriptional activity of SRC-3 in MDA-MB-231 cells transduced with adenoviruses expressing GFP or PFKFB4 cultured in the presence of 5 mM, 15 mM or 25 mM glucose. $n = 6$ (pBIND) and $n = 3$ (pBIND-SRC-3) biological cell samples. *$P < 0.000001$, two-way ANOVA with Tukey's multiple comparisons test. Data in **a**–**c**, **f**–**h** are representative of three biologically independent experiments with similar results, and in **d**, **e** are representative of two biologically independent experiments each run with three different reactions all showing similar results and peptide coverage. Data are mean ± s.d.
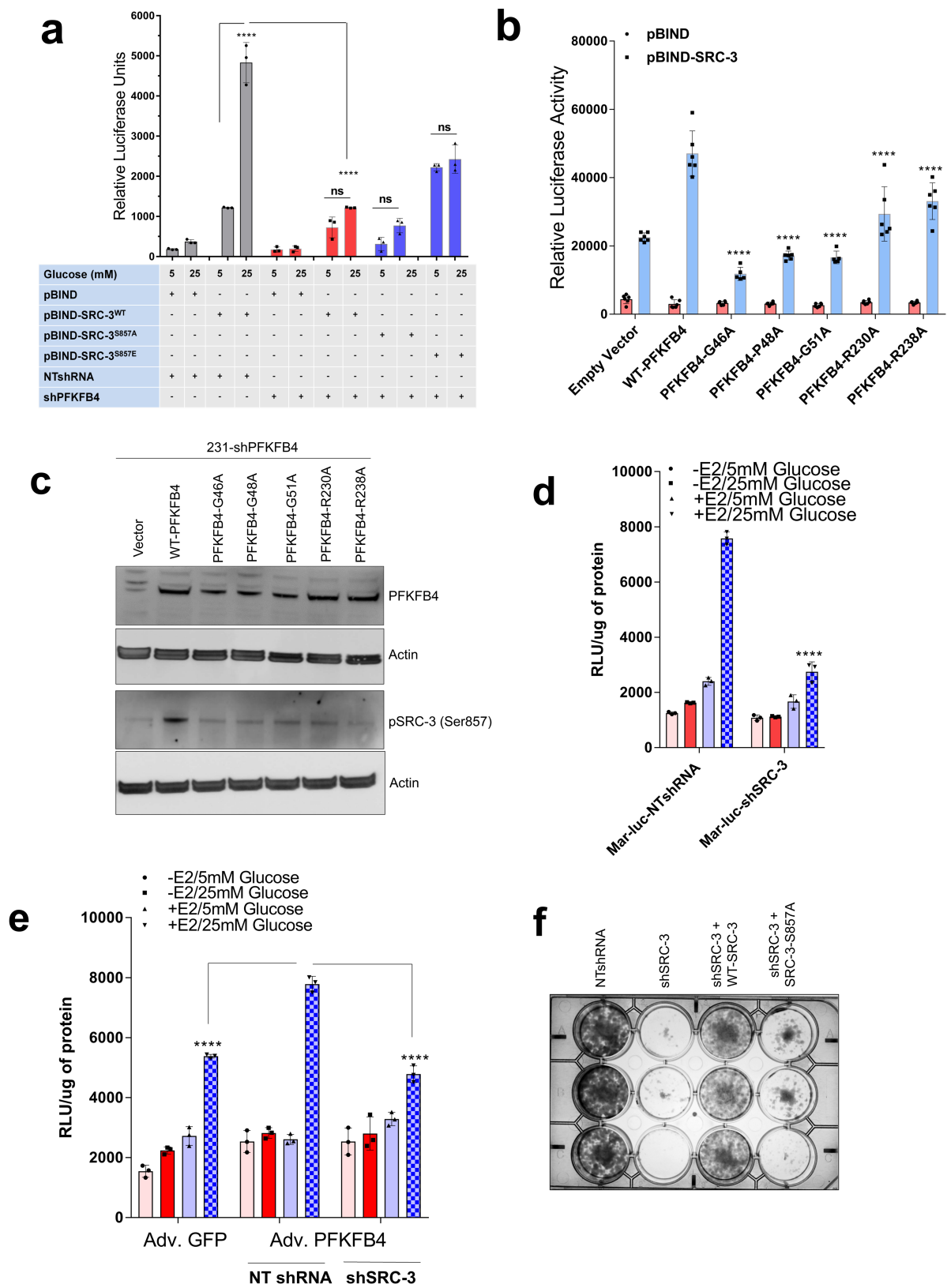
Extended Data Fig. 4 | See next page for caption.

**Extended Data Fig. 4 | Ser857 phosphorylation enhances SRC-3 transcriptional activity. a**, Relative luciferase activity showing the activity of wild-type SRC-3, and the Ser857Ala and Ser857Glu SRC-3 mutants in MDA-MB-231 cells transduced with lentivirus expressing non-targeting shRNA or *PFKFB4* shRNA cultured in the presence of 5 mM or 25 mM glucose. *n* = 3 biological cell samples. *P < 0.000001, two-way ANOVA with Tukey's multiple comparisons test. **b**, Relative luciferase activity showing the activity of SRC-3 in MDA-MB-231 cells stably expressing lentivirus *PFKFB4* shRNA and cultured in the presence of 25 mM glucose. The cells are then co-transfected with empty vector, wild-type PFKFB4 and PFKFB4 mutants Gly46Ala, Pro48Ala, Gly51Ala, Arg230Ala and Arg238Ala. *n* = 6 biological cell samples. *P < 0.000001, two-way ANOVA with Tukey's multiple comparisons test. **c**, MDA-MB-231 cells stably expressing shRNAs targeting *PFKFB4* (231-shPFKFB4) were transfected with constructs expressing empty vector (vector), wild-type PFKFB4, and PFKFB4 mutants Gly46Ala, Pro48Ala, Gly51Ala, Arg230Ala and Arg238Ala, and cultured in presence of 25 mM glucose. Protein levels of pSRC-3-Ser857, PFKFB4 and β-actin were detected by immunoblotting.

d, Relative luciferase activity showing the activity of oestrogen receptor-α (ERα) in MCF7-Mar-luc cells transduced with lentivirus expressing non-targeting shRNA or *SRC-3* shRNA cultured in the presence of 5 mM or 25 mM glucose stimulated with 100 nm E2, or with ethanol control (−E2). *n* = 3 biological cell samples. *P < 0.000001, two-way ANOVA with Tukey's multiple comparisons test. **e**, Relative luciferase activity showing the activity of ERα in MCF7-Mar-luc cells transduced with adenovirus expressing GFP or PFKFB4. Cells transduced with PFKFB4 adenovirus were infected with non-targeting shRNA or *SRC-3* shRNA after 2 days and then cultured in the presence of 5 mM or 25 mM glucose stimulated with ethanol (−E2) or with 100 nM E2. *n* = 3 biological cell samples. *P < 0.000001, two-way ANOVA with Tukey's multiple comparisons test. **f**, Survival assay in MCF7 cells showing the effect of non-targeting shRNA, *SRC-3* shRNA, and re-expression of wild-type SRC-3 or SRC-3(Ser857Ala) mutant in SRC-3-depleted cells cultured in charcoal-stripped medium supplemented with 25 mM glucose and E2 for 7 days. *n* = 3 biological independent data are shown. All data are representative of three independent experiments with similar results, and shown as mean ± s.d.
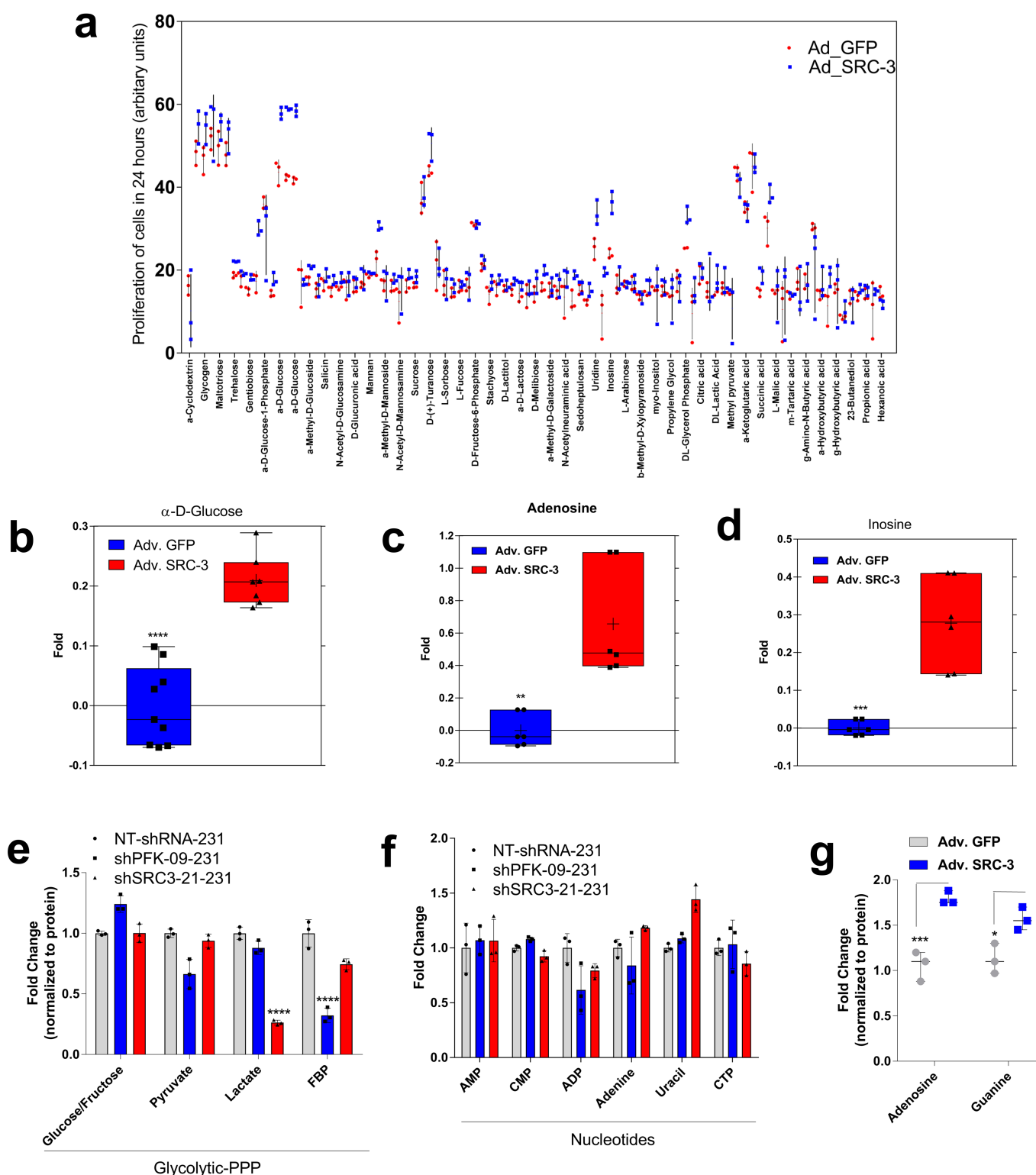
Extended Data Fig. 5 | Increased glucose and purines are required for SRC-3-dependent growth. a, Real-time measurement of MCF10A cell proliferation transduced with adenoviruses expressing GFP or SRC-3 in the presence of 93 different metabolites. $n = 3$ independent plates run for each sample. **b**, Relative growth of MCF10A cells transduced with adenoviruses expressing GFP or SRC-3 in the presence of D-glucose (**b**), adenosine (**c**) and inosine (**d**). $n = 6$ biological cell samples. **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, unpaired t-test two tailed. Boxes are as in Fig. 2e. **e, f**, Relative levels of intermediary metabolites in MDA-MB-231

cells after treatment with shRNAs targeting *PFKFB4* or *SRC-3* compared to control shRNA. **e**, Glycolytic and PPP metabolites. **f**, Nucleotides. $n = 3$ biological independent samples. *$P < 0.05$, two-way ANOVA with Tukey's multiple comparisons test. **g**, Total levels of purines in MCF10A cells transduced with with adenoviruses expressing GFP or SRC-3. $n = 3$ biological independent samples. *$P < 0.05$, ***$P < 0.001$, two-way ANOVA with Tukey's multiple comparisons test. See Source Data for exact P values. Unless stated otherwise, data are mean ± s.d.
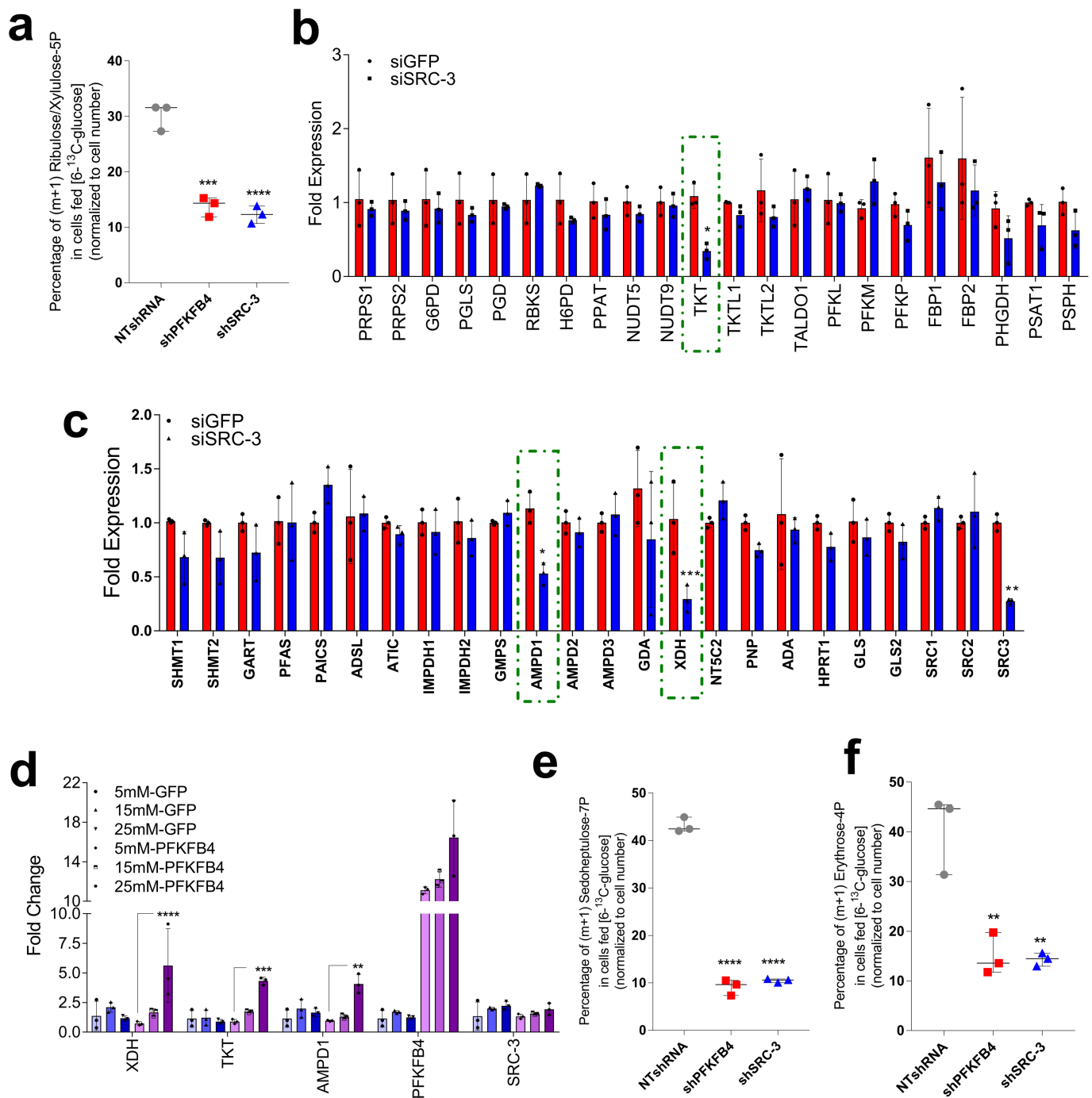
Extended Data Fig. 6 | SRC-3 drives the purine synthesis program under conditions of active glycolysis. a, MDA-MB231 cells stably expressing control shRNA, *PFKFB4* shRNA and *SRC-3* shRNA were fed with [6-^{13}C]glucose. Ribulose/xylulose-5P ($m + 1$) labelling from [6-^{13}C]glucose is shown. $n = 3$ biological cell samples. ***$P = 0.00013$, ****$P = 0.000078$, one-way ANOVA with Tukey's multiple comparisons test. **b**, Genes involved in oxidative and non-oxidative PPP. $n = 3$ biological cell samples. *$P = 0.0431$, two-way ANOVA with Sidak's multiple comparisons test. **c**, Genes involved in nucleotide synthesis. $n = 3$ biological cell samples. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-way ANOVA with Sidak's multiple comparisons test. **d**, mRNA expression of the metabolic enzymes *TKT*, *XDH* and *AMPD1* in MDA-MB-231

cells transduced with adenovirus expressing GFP (control) and PFKFB4 cultured in the presence of 5 mM, 15 mM or 25 mM glucose. $n = 3$ biological cell samples. **$P < 0.01$, ***$P < 0.001$, ****$P = 0.0001$, two-way ANOVA with Dunnett's multiple comparisons test. **e**, **f**, MDA-MB231 cells stably expressing control shRNA, *PFKFB4* shRNA and *SRC-3* shRNA were fed with [6-^{13}C]glucose. Seduheptulose-7P ($m + 1$) (**e**) and erythrose-4P (**f**) labelling from [6-^{13}C]glucose are shown. $n = 3$ biological cell samples. **$P < 0.01$, ****$P = 0.001$, two-way ANOVA with Dunnett's multiple comparisons test (**e**) or with Tukey's multiple comparison test (**f**). Boxes are as in Fig. 2e. See Source Data for exact P values. Unless stated otherwise, data are mean ± s.d.
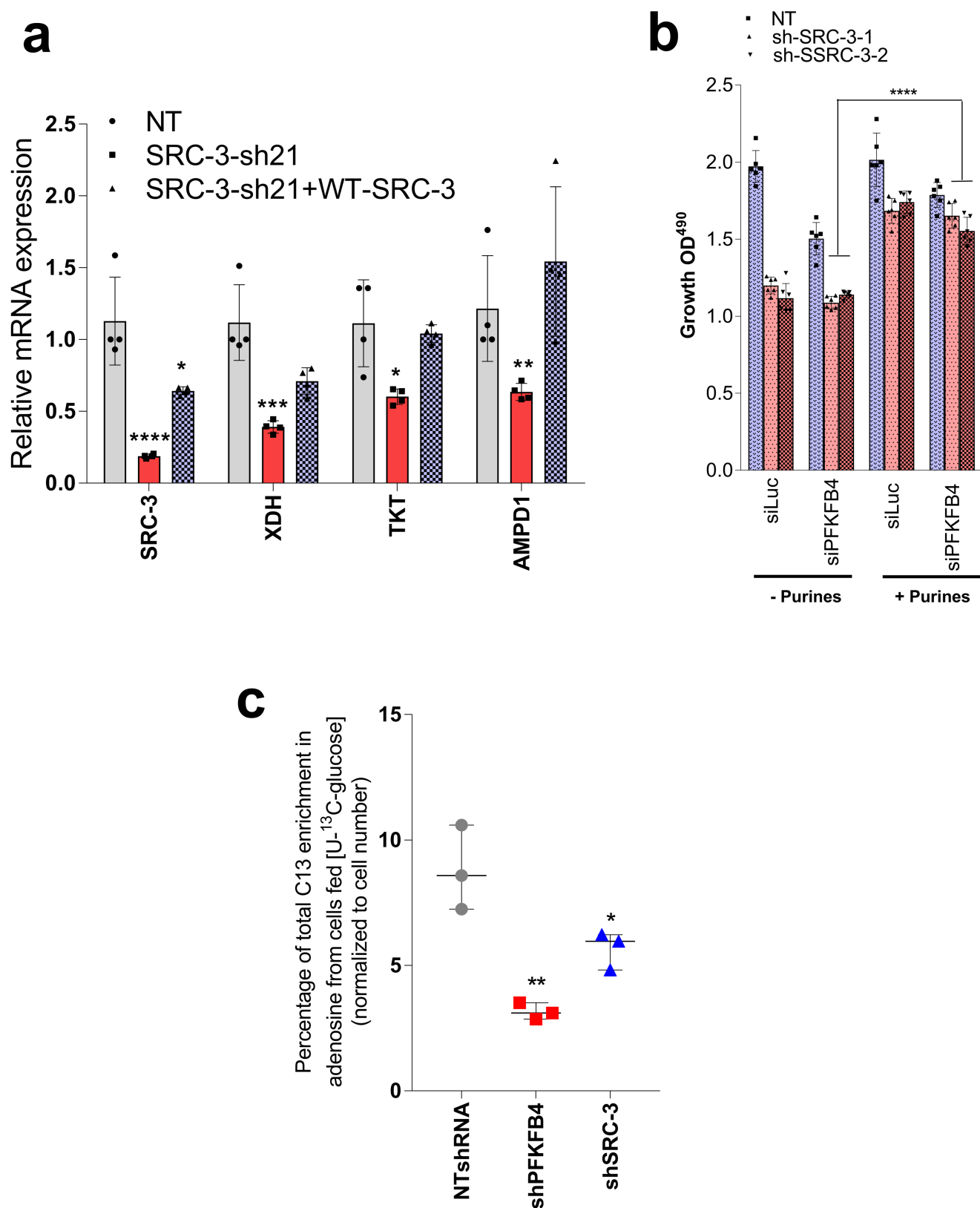
Extended Data Fig. 7 | Growth defect due to loss of SRC-3 or PFKFB4 is rescued by exogenous purines. a, Expression of the metabolic enzymes encoded by *TKT*, *XDH*, *AMPD1* and *SRC-3* in MDA-MB-231 cells expressing control shRNA, *SRC-3* shRNA or *SRC-3* shRNA plus re-expression of shRNA-resistant wild-type SRC-3 protein (shSRC-3-21 + WT-SRC-3). $n = 4$ biological cell samples. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P = 0.0001$, two-way ANOVA with Tukey's multiple comparisons test. **b**, Relative proliferation of MDA-MB-231 cells expressing shRNA targeting *SRC-3* (shSRC-3#01 and shSRC-3#02) or non-targeting control shRNA after treatment with siRNAs targeting luciferase (siLuc; as a control) or *PFKFB4*

under the conditions indicated. $n = 6$ samples from biologically independent experiments. $****P < 0.000001$, two-way ANOVA with Tukey's multiple comparisons test. **c**, MDA-MB231 cells stably expressing control shRNA, *PFKFB4* shRNA or *SRC-3* shRNA were fed with [U-^{13}C]glucose for 48 h. Adenosine ^{13}C-labelling from [U-^{13}C]glucose is shown. $n = 3$ samples from biologically independent experiments. one-way ANOVA with Tukey's multiple comparisons test. Boxes are as in Fig. 2e. Data are representative of three biologically independent experiments with similar results. See Source Data for exact P values. Unless stated otherwise, data are mean ± s.d.

Extended Data Fig. 8 | See next page for caption.
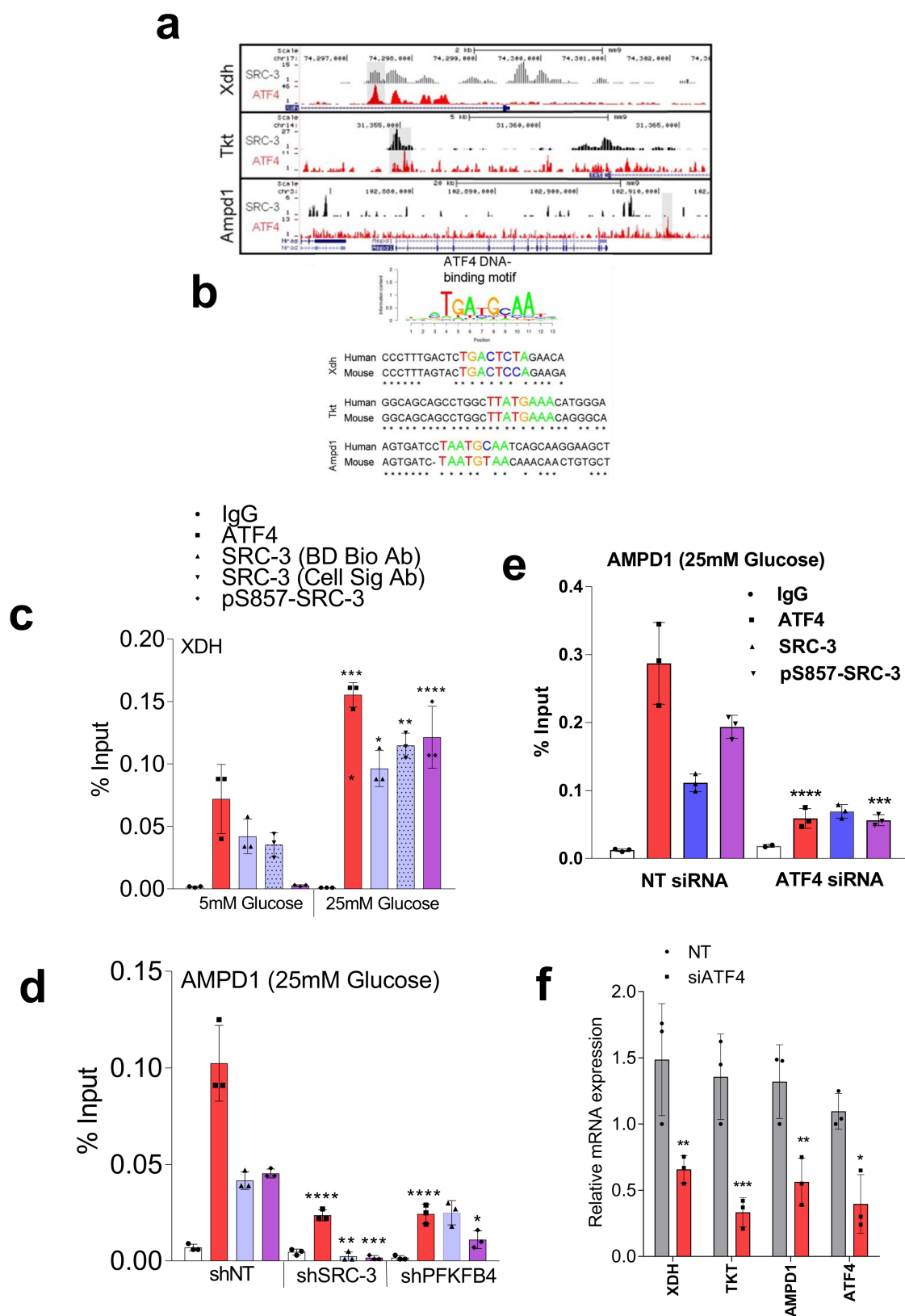
Extended Data Fig. 8 | PFKFB4–SRC-3 stabilizes ATF4 transcription factor to promote purine synthesis. a, Chromatin localization peaks of SRC-3 and ATF4 on *Tkt*, *Xdh* and *Ampd1* genes in mouse liver. **b**, ATF4-binding peaks are conserved on three SRC-3 target purine biosynthetic genes in both mouse and human genomes. **c**, Chromatin immunoprecipitation (ChIP) of ATF4, total SRC-3 and pSRC-3-Ser857 from MDA-MB-231 cells treated with 5 mM or 25 mM glucose compared to an IgG isotype control. qPCR was performed to determine amount of promoter enrichment. **d**, ChIP–qPCR was performed from MDA-MB-231 cells cultured in 25 mM glucose expressing *SRC-3* shRNA, *PFKFB4* shRNA or control shRNA. $n = 3$ biological cell samples. $*P < 0.01$, $**P < 0.0001$, $***P < 0.00005$, $****P < 0.000001$, one-way ANOVA with Tukey's

multiple comparisons test compared to 5 mM glucose groups (**c**) and compared to NT shRNA group (**d**). **e**, ChIP of ATF4, total SRC-3 (BD Biosciences antibody), and pSRC-3-Ser857 from MDA-MB-231 cells on the *AMPD1* promoter treated with non-targeting siRNA or siRNA against *ATF4*, and cultured in presence of 25 mM glucose compared to an IgG isotype control. qPCR was performed to determine the amount of promoter enrichment. $n = 3$ biological cell samples. $***P < 0.001$, $****P < 0.000001$, one-way ANOVA with Tukey's multiple comparisons test. **f**, mRNA expression of *TKT*, *XDH*, *AMPD1* and *SRC-3* in MDA-MB-231 cells expressing siRNA targeting control or *ATF4* siRNA. $n = 3$ biological cell samples. two-way ANOVA with Sidak's multiple comparisons test. See Source Data for exact *P* values. Data are mean ± s.d.

a

NTshRNA

shSRC-3

shPFKFB4

231-shSRC-3
+ WT-SRC-3

231-shSRC-3
+ S857A

b

231-NT   231-shSRC-3   231-shPFKFB4

c

Ki67 Positive Cells/Field

- 231-shNT
- 231-shSRC-3
- 231-shPFKFB4

****
****

231-shNT   231-shSRC-3   231-shPFKFB4

d

Tumor Diameter (cm)

- NTshRNA
- 231-shSRC-3-21
- 231-shPFKFB4-20
- 231-shSRC-3-21+SRC-3
- 231-shSRC-3-21+S857A

****
****
****

Time (days)

e

NTshRNA   shSRC-3   shSRC-3 +WT   shSRC-3 +S857A

SRC-3

1.0   0.05   0.5   0.75

Actin

f

Photon Flux

SRC-3 WT   SRC-3-S857A   shSRC-3   shPFKFB4

g

Log Change over NTshRNA

- NTshRNA
- shSRC-3
- shPFKFB4

****   ****   ***   ****   ****   **   ***   ****   ****

SRC-3   PFKFB4   TKT   XDH   AMPD1

h

Expression of PFKFB4 in TCGA Breast dataset

mRNA expression

P-value = 1.54e-09   P-value = 5.22-09   P-value = 5.96e-33   P-value = 8.50e-25

Basal_Normal   Basal   Normal_HER2   HER2   Normal_Luminal A   Luminal A   Normal_Luminal B   Luminal B

Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | The PFKFB4–SRC-3 axis promotes breast tumour growth and metastasis. a, Primary tumours resected out after 6 weeks. **b**, Ki67 staining of primary tumours from animals injected with MDA-MB-231 cells stably expressing control shRNA, *SRC-3* shRNA or *PFKFB4* shRNA. Data are representative of five fields per slide from $n = 5$ animals per group with similar findings. Scale bar, 100 μm. **c**, Quantification of Ki67-positive cells in the tumour. $n = 5$ animals per group, average of five fields counted from each slide. ****$P = 0.0001$, one-way ANOVA with Dunnett's multiple comparisons test. **d**, Primary tumour growth in animals injected with MDA-MB-231 cells stably expressing shRNA targeting *SRC-3*, *PFKFB4*, or expression of wild-type SRC-3 or the Ser857Ala mutant in the SRC-3-depleted cells. $n = 5$ animals per group. *$P < 0.000001$, two-way ANOVA with Tukey's multiple comparisons test. **e**, Immunoblot showing the relative expression of SRC-3 in primary tumours from MDA-MB-231 cells stably expressing control shRNA, *SRC-3* shRNA, or after re-expression of wild-type SRC-3 or the Ser857Ala mutant in the SRC-3-depleted cells. $n = 5$ animals per group was pooled to generate the tumour lysate used for analysis. **f**, Graph representing the photon flux of animals from different groups. $n = 5$ animals for wild-type SRC-3, the Ser857Ala mutant and *PFKFB4* shRNA, and $n = 4$ animals for *SRC-3* shRNA. *$P < 0.0001$, one-way ANOVA with Dunnett's multiple comparisons test. Line shows median with range. **g**, mRNA expression of three metabolic enzymes (*TKT*, *XDH* and *AMPD1*), *SRC-3* and *PFKFB4* from the primary tumours. $n = 5$ animals per group. **$P < 0.05$, ***$P < 0.001$, ****$P = 0.0001$, two-way ANOVA with Tukey's multiple comparisons test. **h**, Expression of *PFKFB4* in patients with breast cancer across different subtypes compared to normal breast tissue. Normal basal $= 17$; basal $= 139$; normal_Her2 $= 9$; Her2 $= 67$; normal luminal A $= 62$; luminal A $= 418$; normal luminal B $= 21$ and LumB $= 186$. Line in the centre of the rectangle represents the median, top edge of the rectangle represents the third quartile, bottom edge of the rectangle represents the first quartile, top whisker represents the maximum and bottom whisker represents the minimum. All data are representative of three biologically independent experiments with similar results, and are shown as mean $\pm$ s.d. unless otherwise stated. See Source Data for exact *P* values.
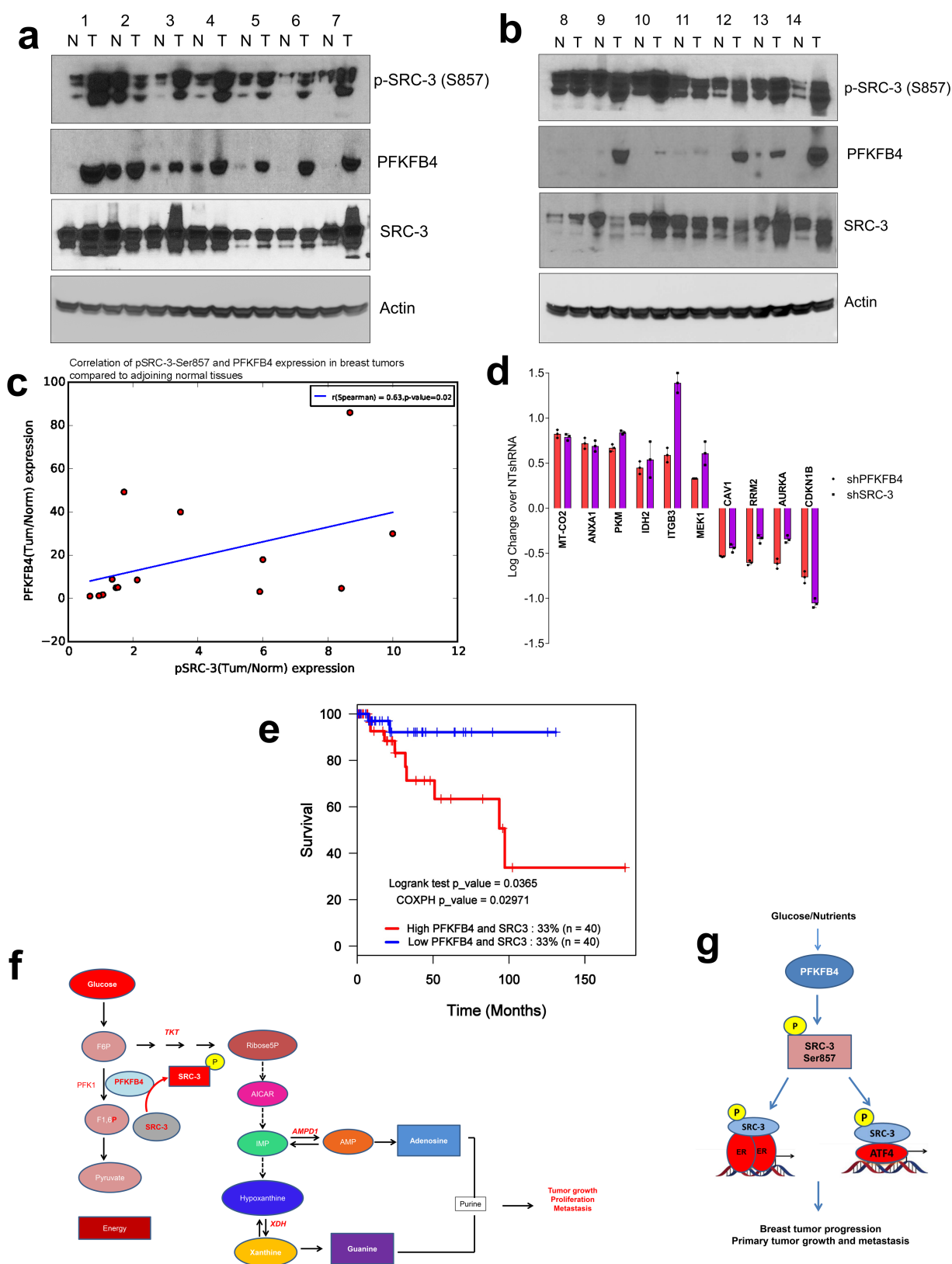
Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The PFKFB4–SRC-3 axis drives transcriptional programming in patients with breast cancer. a, **b**, Expression of pSRC-3, SRC-3 and PFKFB4 in ER$^+$ breast tumour specimens and matched adjoining normal tissues as detected by immunoblotting. $n = 14$ patients with ER$^+$ breast cancer. **c**, Semi-quantitative levels of bands shown in **a** and **b**, analysed by densitometry using UVP Vision Works LS software, and normalized relative to actin to calculate the fold change (tumour/normal) and plotted to obtain the correlation between PFKFB4 and pSRC-3-Ser857 expression. $n = 14$ normal and tumour tissues. $R = 0.63$, $P = 0.02$ Spearman's rank correlation coefficient. **d**, log fold change in protein expression of the PFKFB4–SRC-3 signature compared to the control knockdown (non-targeting shRNA) as determined using a parametric t-test as implemented in the python (spicy) statistical system. Significance $P < 0.05$ and fold change exceeding $1.25\times$ were used to classify true regulators of SRC-3 activity. $n = 3$ biologically independent samples. **e**, Kaplan–Meier survival plot showing poor survival of patients with breast cancer with basal subtype (triple-negative) disease exhibiting an increased expression of a common proteomic signature induced by the PFKFB4 and SRC-3 axis. The cohort of patients was collected by the TCGA. $P = 0.0365$, log-rank test; $P = 0.02971$, Cox proportional hazards, two-sided. **f**, Cartoon model describing the crosstalk between glycolysis and purine generation highlighting the essential steps regulated by pSRC-3-Ser857. This PFKFB4-dependent SRC-3 phosphorylation enhances mRNA expression of genes involved in purine metabolism driving breast tumour growth, proliferation and metastasis. AICAR, 5-aminoimidazole-4-carboxamide ribonucleotide; AMP, adenosine monophosphate; F1,6-P, fructose 1,6 bisphosphate; IMP, inosine monophosphate. **g**, Model showing that, in glycolytic breast tumours, activated PFKFB4 drives SRC-3 phosphorylation at Ser857, which then activates ER-positive primary tumour growth in conjunction with E2-liganded ER, as well in ER-negative/recurrent tumours in conjunction with ATF4, driving aggressive metastatic disease. Data are mean ± s.d.

LETTER

Fatal swine acute diarrhoea syndrome caused by an HKU2–related coronavirus of bat origin

Peng Zhou[1,12], Hang Fan[2,12], Tian Lan[3,4,12], Xing-Lou Yang[1], Wei-Feng Shi[5], Wei Zhang[1], Yan Zhu[1], Ya-Wei Zhang[2], Qing-Mei Xie[3,4], Shailendra Mani[6], Xiao-Shuang Zheng[1], Bei Li[1], Jin-Man Li[2], Hua Guo[1], Guang-Qian Pei[2], Xiao-Ping An[2], Jun-Wei Chen[3,4], Ling Zhou[3,4], Kai-Jie Mai[3,4], Zi-Xian Wu[3,4], Di Li[3,4], Danielle E. Anderson[6], Li-Biao Zhang[7], Shi-Yue Li[8], Zhi-Qiang Mi[2], Tong-Tong He[2], Feng Cong[9], Peng-Ju Guo[9], Ren Huang[9], Yun Luo[1], Xiang-Ling Liu[1], Jing Chen[1], Yong Huang[2], Qiang Sun[2], Xiang-Li-Lan Zhang[2], Yuan-Yuan Wang[2], Shao-Zhen Xing[2], Yan-Shan Chen[3,4], Yuan Sun[3,4], Juan Li[5], Peter Daszak[10]*, Lin-Fa Wang[6]*, Zheng-Li Shi[1]*, Yi-Gang Tong[2,11]* & Jing-Yun Ma[3,4]*

Cross-species transmission of viruses from wildlife animal reservoirs poses a marked threat to human and animal health[1]. Bats have been recognized as one of the most important reservoirs for emerging viruses and the transmission of a coronavirus that originated in bats to humans via intermediate hosts was responsible for the high-impact emerging zoonosis, severe acute respiratory syndrome (SARS)[2–10]. Here we provide virological, epidemiological, evolutionary and experimental evidence that a novel HKU2-related bat coronavirus, swine acute diarrhoea syndrome coronavirus (SADS-CoV), is the aetiological agent that was responsible for a large-scale outbreak of fatal disease in pigs in China that has caused the death of 24,693 piglets across four farms. Notably, the outbreak began in Guangdong province in the vicinity of the origin of the SARS pandemic. Furthermore, we identified SADS-related CoVs with 96–98% sequence identity in 9.8% (58 out of 591) of anal swabs collected from bats in Guangdong province during 2013–2016, predominantly in horseshoe bats (*Rhinolophus* spp.) that are known reservoirs of SARS-related CoVs. We found that there were striking similarities between the SADS and SARS outbreaks in geographical, temporal, ecological and aetiological settings. This study highlights the importance of identifying coronavirus diversity and distribution in bats to mitigate future outbreaks that could threaten livestock, public health and economic growth.

The emergence of SARS in southern China in 2002, which was caused by a previously unknown coronavirus (SARS-CoV)[11–15] and has led to more than 8,000 human infections and 774 deaths (http://www.who.int/csr/sars/en/), highlights two new frontiers in emerging infectious diseases. First, it demonstrates that coronaviruses are capable of causing fatal diseases in humans. Second, the identification of bats as the reservoir for SARS-related coronaviruses, and the fact that SARS-CoV[3–10] probably originated in bats, firmly establishes that bats are an important source of highly lethal zoonotic viruses, such as Hendra, Nipah, Ebola and Marburg viruses[16].

Here we report on a series of fatal swine disease outbreaks in Guangdong province, China, approximately 100 km from the location of the purported index case of SARS. Most strikingly, we found that the causative agent of this swine acute diarrhoea syndrome (SADS) is a novel HKU2-related coronavirus that is 98.48% identical in genome sequence to a bat coronavirus, which we detected in 2016 in bats in a cave in the vicinity of the index pig farm. This new virus (SADS-CoV) originated from the same genus of horseshoe bats (*Rhinolophus*) as SARS-CoV.

From 28 October 2016 onwards, a fatal swine disease outbreak was observed in a pig farm in Qingyuan, Guangdong province, China, very close to the location of the first known index case of SARS in 2002, who lived in Foshan (Extended Data Fig. 1a). Porcine epidemic diarrhoea virus (PEDV, a coronavirus) had caused prior outbreaks at this farm, and was detected in the intestines of deceased piglets at the start of the outbreak. However, PEDV could no longer be detected in deceased piglets after 12 January 2017, despite accelerating mortality (Fig. 1a), and extensive testing for other common swine viruses yielded no results (Extended Data Table 1). These findings suggested that this was an outbreak of a novel disease. Clinical signs are similar to those caused by other known swine enteric coronaviruses[17,18] and include severe and acute diarrhoea and acute vomiting, leading to death due to rapid weight loss in newborn piglets that are less than five days of age. Infected piglets died 2–6 days after disease onset, whereas infected sows suffered only mild diarrhoea and most sows recovered within two days. The disease caused no signs of febrile illness in piglets or sows. The mortality rate was as high as 90% in piglets that were five days or younger, whereas in piglets that were older than eight days, the mortality dropped to 5%. Subsequently, SADS-related outbreaks were found in three additional pig farms within 20–150 km of the index farm (Extended Data Fig. 1a) and, by 2 May 2017, the disease had caused the death of 24,693 piglets at these four farms (Fig. 1a). In farm A alone, 64% (4,659 out of 7,268) of all piglets that were born in February died. The outbreak has abated, and measures that were taken to control SADS included separation of sick sows and piglets from the rest of the herd. A qPCR test described below was used as the main diagnostic tool to confirm SADS-CoV infection.

A sample collected from the small intestine of a diseased piglet was analysed by metagenomics analysis using next-generation sequencing (NGS) to identify potential aetiological agents. Of the 15,256,565 total reads obtained, 4,225 matched sequences of the bat CoV HKU2, which was first detected in Chinese horseshoe bats in Hong Kong and Guangdong province, China[19]. By de novo assembly and targeted PCR, we obtained a 27,173-bp CoV genome that shared 95% sequence identity to HKU2-CoV (GenBank accession number NC_009988). Thirty-three full genome sequences of SADS-CoV were subsequently obtained (8 from farm A, 5 from farm B, 11 from farm C and 9 from farm D) that were 99.9% identical to each other (Supplementary Table 1).
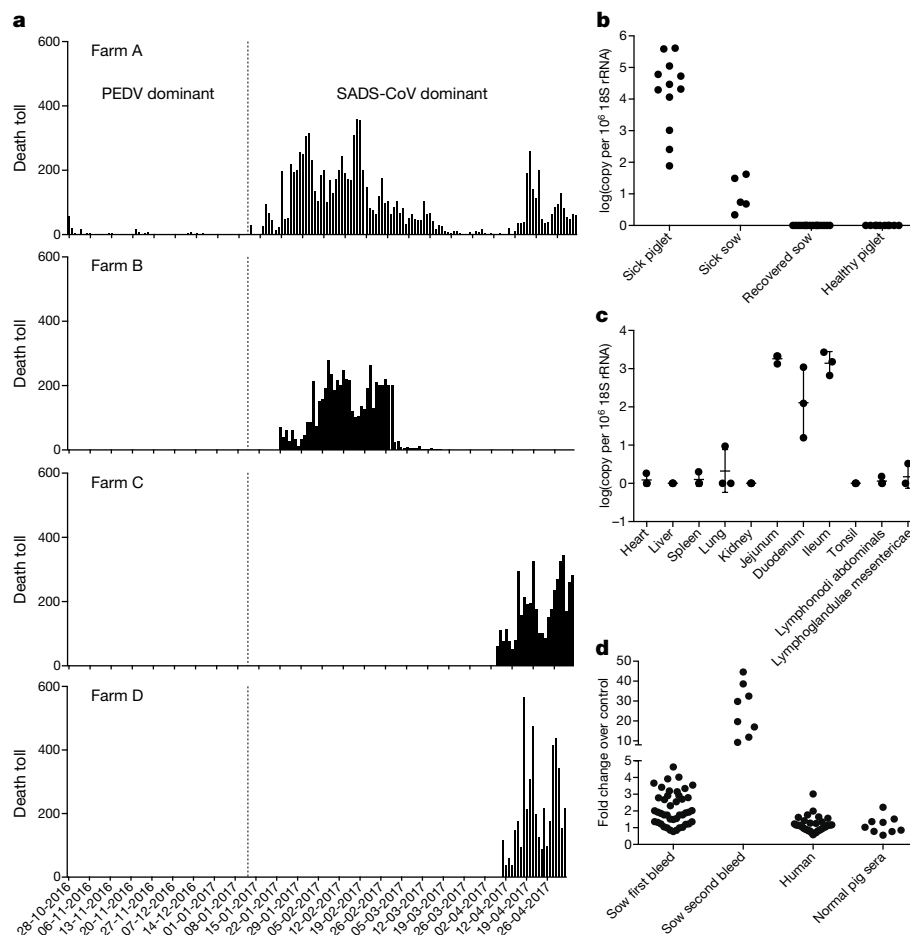
Fig. 1 | Detection of SADS-CoV infection in pigs in Guangdong, China.
a, Records of daily death toll on the four farms from 28 October 2016 to 2 May 2017. **b**, Detection of SADS-CoV by qPCR. The y axis shows the log(copy number per 10^6 copies of 18S rRNA). $n = 12$ sick piglets, 5 sick sows, 16 recovered sows and 10 healthy piglets. **c**, Tissue distribution of SADS-CoV in diseased pigs. $n = 3$. Data are mean ± s.d.; dots represent individual values. **d**, Detection of SADS-CoV antibodies. $n = 46$ sows from whom serum was first taken in the first three weeks of the outbreak (First bleed), $n = 8$ sows from whom serum was taken again (Second bleed) at more than one month after the onset of the outbreak, $n = 8$ sera from healthy pig controls, $n = 35$ human sera from pig farmers.

Using qPCR targeting the nucleocapsid gene (Supplementary Table 2), we detected SADS-CoV in acutely sick piglets and sows, but not in recovered or healthy pigs on the four farms, nor in nearby farms that showed no evidence of SADS. The virus replicated to higher titres in piglets than in sows (Fig. 1b). SADS-CoV displayed tissue tropism of the small intestine (Fig. 1c), as observed for other swine enteric coronaviruses[20]. Retrospective PCR analysis revealed that SADS-CoV was present on farm A during the PEDV epidemic, where the first strongly positive SADS-CoV sample was detected on 6 December 2016. From mid-January onwards, SADS-CoV was the dominant viral agent detected in diseased animals (Extended Data Fig. 1b). It is possible that the presence of PEDV early in the SADS-CoV outbreak may have somehow facilitated or enhanced spillover and amplification of SADS. However the fact that the vast majority of piglet mortality occurred after PEDV infection had become undetectable suggests that SADS-CoV itself causes a lethal infection in pigs that was responsible for these large-scale outbreaks, and that PEDV does not directly contribute to its severity in individual pigs. This was supported by the absence of PEDV and other known swine diarrhoea viruses during the peak and later phases of the SADS outbreaks in the four farms (Extended Data Table 1).

We rapidly developed an antibody assay based on the S1 domain of the spike (S) protein using a luciferase immunoprecipitation system[21]. Because SADS occurs acutely and has a rapid onset in piglets, serological investigation was conducted only in sows. Among 46 recovered sows tested, 12 were seropositive for SADS-CoV within three weeks

of infection (Fig. 1d). To investigate possible zoonotic transmission, serum samples from 35 farm workers who had close contact with sick pigs were also analysed using the same luciferase immunoprecipitation system approach and none were positive for SADS-CoV.

Although the overall genome identity of SADS-CoV and HKU2-CoV is 95%, the S gene sequence identity is only 86%, suggesting that the previously reported HKU2-CoV is not the direct progenitor of SADS-CoV, but that they may have originated from a common ancestor. To test this hypothesis, we developed a SADS-CoV-specific qPCR assay based on its RNA-dependent RNA polymerase (*RdRp*) gene (Supplementary Table 2) and screened 591 bat anal swabs collected between 2013 and 2016 from seven different locations in Guangdong province (Extended Data Fig. 1a). A total of 58 samples (9.8%) tested positive (Extended Data Table 2), all were from *Rhinolophus* spp. bats that are also the natural reservoir hosts of SARS-related coronaviruses[3–10]. Four complete genome sequences with the highest *RdRp* PCR-fragment sequence identity to that of SADS-CoV were determined by NGS. They are very similar in size (27.2 kb) compared to SADS-CoV (Fig. 2a) and we tentatively call them SADS-related coronaviruses (SADSr-CoV). Overall sequence identity between SADSr-CoV and SADS-CoV ranges from 96 to 98%. Most importantly, the S protein of SADS-CoV shared more than 98% sequence identity with sequences of two of the SADSr-CoVs (samples 162149 and 141388), compared to 86% with HKU2-CoV. The major sequence differences among the four SADSr-CoV genomes were found in the predicted coding regions of the *S* and *NS7a* and *NS7b* genes (Fig. 2a). In addition, the coding region
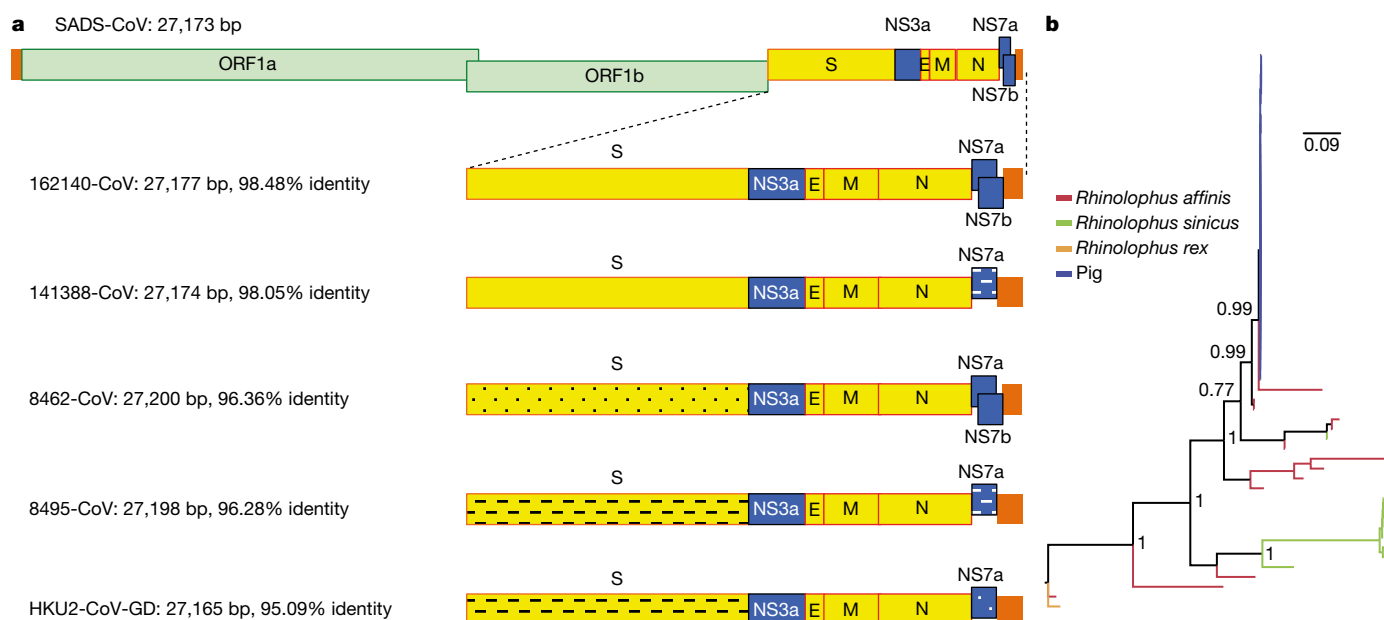
Fig. 2 | Genome and phylogenetic analysis of SADS-CoV and SADSr-CoV. a, Genome organization and comparison. Colour-coding for different genomic regions as follows. Green, non-structural polyproteins *ORF1a* and *ORF1b*; yellow, structural proteins *S*, *E*, *M* and *N*; blue, accessory proteins *NS3a*, *NS7a* and *NS7b*; Orange, untranslated regions. The level of sequence identity of SADSr-CoV to SADS-CoV is illustrated by different patterns of boxes: Solid colour, highly similar; Dotted fill, moderately similar; Dashed fill, least similar. **b**, Phylogenetic analysis of 57 *S1* sequences (33 from SADS-CoV and 24 from SADSr-CoV). Different colours represent different host species as shown on the left. Scale bar, nucleotide substitutions per site.

of the S protein N-terminal (S1) domain was determined from 19 bat SADSr-CoVs to enable more detailed phylogenetic analysis.

The phylogeny of S1 and the full-length genome revealed a high genetic diversity of alphacoronaviruses among bats and strong coevolutionary relationships with their hosts (Fig. 2b and Extended Data Fig. 2), and showed that SADS-CoVs were more closely related to SADSr-CoVs from *Rhinolophus affinis* than from *Rhinolophus sinicus*, in which HKU2-CoV was found. Both phylogenetic and haplotype network analyses demonstrated that the viruses from the four farms probably originated from their reservoir hosts independently (Extended Data Fig. 3), and that a few viruses might have undergone further genetic recombination (Extended Data Fig. 4). However, molecular clock analysis of the 33 SADS-CoV genome sequences failed to establish a positive association between sequence divergence and sampling date. Therefore, we speculate that either the virus was introduced into pigs from bats multiple times, or that the virus was introduced into pigs once, but subsequent genetic recombination disturbed the molecular clock.

For viral isolation, we tried to culture the virus in a variety of cell lines (see Methods for details) using intestinal tissue homogenates as starting material. Cytopathogenic effects were observed in Vero cells only after five passages (Extended Data Fig. 5a, b). The identity of SADS-CoV was verified in Vero cells by immunofluorescence microscopy (Extended Data Fig. 5c, d) and by whole-genome sequencing (GenBank accession number MG557844). Similar results were obtained by other groups[22,23].

Known coronavirus host cell receptors include angiotensin-converting enzyme 2 (ACE2) for SARS-related CoV, aminopeptidase N (APN) for certain alphacoronaviruses, such as human (H)CoV-229E, and dipeptidyl peptidase 4 (DPP4) for Middle East respiratory syndrome (MERS)-CoV[24–26]. To investigate the receptor usage of SADS-CoV, we tested live or pseudotyped SADS-CoV infection on HeLa cells that expressed each of the three molecules. Whereas the positive control worked for SARS-related CoV and MERS-CoV pseudoviruses, we found no evidence of enhanced infection or entry for SADS-CoV, suggesting that none of these receptors functions as a receptor for virus entry for SADS-CoV (Extended Data Table 3).

To fulfill Koch's postulates for SADS-CoV, two different types of animal challenge experiments were conducted (see Methods for

details). The first challenge experiment was conducted with specific pathogen-free piglets that were infected with a tissue homogenate of SADS-CoV-positive intestines. Two days after infection, 3 out of 7 animals died in the challenge group whereas 4 out of 5 survived in the control group. Incidentally, the one piglet that died in the control group was the only individual that did not receive colostrum due to a shortage in the supply. It is thus highly likely that lack of nursing and
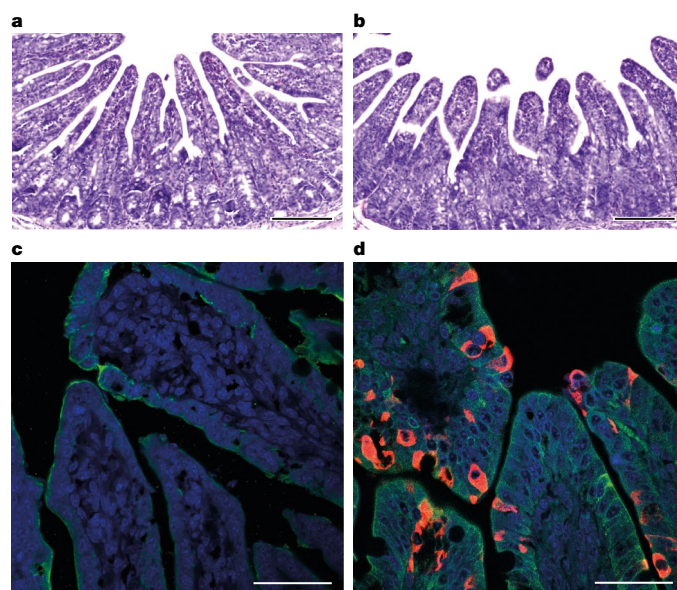
Fig. 3 | Immunohistopathology of SADS-CoV infected tissues. a–d, Sections of jejunum tissue from control (**a**, **c**) and infected (**b**, **d**) farm piglets four days after inoculation were stained with haematoxylin and eosin (**a**, **b**) or rabbit anti-SADSr-CoV N serum (red), DAPI (blue) and mouse antibodies against epithelial cell markers cytokeratin 8, 18 and 19 (green) in (**c**, **d**). SADS-CoV N protein is evident in epithelial cells and deeper in the tissue of infected piglets, which exhibit villus shortening. Scale bars, 200 μm (**a**, **b**) and 50 μm (**c**, **d**). The experiment was conducted three times independently with similar results.

inability to access colostrum was responsible for the death (Extended Data Table 4). For the second challenge, healthy piglets were acquired from a farm in Guangdong that had been free of diarrheal disease for a number of weeks before the experiment, and were infected with the cultured isolate of SADS-CoV or tissue-culture medium as control. Of those inoculated with SADS-CoV, 50% (3 out of 6) died between 2 and 4 days after infection, whereas all control animals survived (Extended Data Table 5). All animals in the infected group suffered watery diarrhoea, rapid weight loss and intestinal lesions (determined after euthanasia upon experiment termination, Extended Data Tables 4, 5). Histopathological examination revealed marked villus atrophy in SADS-CoV inoculated farm piglets four days after inoculation but not in control piglets (Fig. 3a, b) and viral N protein-specific staining was observed mainly in small intestine epithelial cells of the inoculated piglets (Fig. 3c, d).

The current study highlights the value of proactive viral discovery in wildlife, and targeted surveillance in response to an emerging infectious disease event, as well as the disproportionate importance of bats as reservoirs of viruses that threaten veterinary and public health[1]. It also demonstrates that by using modern technological platforms, such as NGS, luciferase immunoprecipitation system serology and phylogenetic analysis, key experiments that traditionally rely on the isolation of live virus can be performed rapidly before virus isolation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0004-7.

1. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
2. Guan, Y. et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
3. Lau, S. K. et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl Acad. Sci. USA* **102**, 14040–14045 (2005).
4. Li, W. et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
5. Ge, X. Y. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
6. He, B. et al. Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China. *J. Virol.* **88**, 7070–7082 (2014).
7. Yang, X. L. et al. Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of severe acute respiratory syndrome coronavirus. *J. Virol.* **90**, 3253–3256 (2016).
8. Wu, Z. et al. ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis.* **213**, 579–583 (2016).
9. Wang, L. et al. Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. *Emerg. Microbes Infect.* **6**, e14 (2017).
10. Hu, B. et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
11. Drosten, C. et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
12. Ksiazek, T. G. et al. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
13. Marra, M. A. et al. The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404 (2003).
14. Peiris, J. S. et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
15. Rota, P. A. et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399 (2003).
16. Wang, L.-F. & Cowled, C. (eds) *Bats and Viruses: A New Frontier of Emerging Infectious Diseases* 1st edn (John Wiley & Sons, Hoboken, 2015).
17. Dong, N. et al. Porcine deltacoronavirus in mainland China. *Emerg. Infect. Dis.* **21**, 2254–2255 (2015).
18. Sun, D., Wang, X., Wei, S., Chen, J. & Feng, L. Epidemiology and vaccine of porcine epidemic diarrhea virus in China: a mini-review. *J. Vet. Med. Sci.* **78**, 355–363 (2016).
19. Lau, S. K. et al. Complete genome sequence of bat coronavirus HKU2 from Chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome. *Virology* **367**, 428–439 (2007).
20. Chen, J. et al. Molecular epidemiology of porcine epidemic diarrhea virus in China. *Arch. Virol.* **155**, 1471–1476 (2010).
21. Burbelo, P. D. et al. Serological diagnosis of human herpes simplex virus type 1 and 2 infections by luciferase immunoprecipitation system assay. *Clin. Vaccine Immunol.* **16**, 366–371 (2009).
22. Gong, L. et al. A new bat-HKU2-like coronavirus in swine, China, 2017. *Emerg. Infect. Dis.* **23**, 1607–1609 (2017).
23. Pan, Y. et al. Discovery of a novel swine enteric alphacoronavirus (SeACoV) in southern China. *Vet. Microbiol.* **211**, 15–21 (2017).
24. Li, W. et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
25. Masters, P. S. & Perlman, S. in *Fields Virology* Vol. 2 (eds Knipe, D. M. & Howley, P. M.) 825–858 (Lippincott Williams & Wilkins, Philadelphia, 2013).
26. Raj, V. S. et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* **495**, 251–254 (2013).

Author contributions L.-F.W., Z.-L.S., P.Z., Y.-G.T., and J.-Y.M. conceived the study. P.Z., W.Z., Y.Z., S.M., X.-S.Z., B.L., X.-L.Y., H.G., D.E.A., Y.L., X.L.L. and J.C. performed qPCR, serology and histology experiments and cultured the virus. H.F., Y.-W.Z., J.-M.L., G.-Q.P., X.-P.A., Z.-Q.M., T.-T.H., Y.H., Q.S., Y.-Y.W., S.-Z.X., X.-L.-L.Z., W.-F.S. and J.L. performed genome sequencing and annotations. T.L., Q.-M.X., J.-W.C., L.Z., K.-J.M., Z.-X.W., Y.-S.C., D.L., Y.S., F.C., P.-J.G. and R.H. prepared the samples and carried out animal challenge experiments. Z.-L.S., P.D., L.-B.Z., S.-Y.L. coordinated collection of bat samples. P.Z., L.-F.W., Z.-L.S. and P.D. had a major role in the preparation of the manuscript.

Competing interests The authors declare no competing interests.

METHODS

Sample collection. Bats were captured and sampled in their natural habitat in Guangdong province (Extended Data Fig. 1) as described previously[4]. Faecal swab samples were collected in viral transport medium (VTM) composed of Hank's balanced salt solution at pH 7.4 containing BSA (1%), amphotericin (15 μg ml^{-1}), penicillin G (100 units ml^{-1}) and streptomycin (50 μg ml^{-1}). Stool samples from sick pigs were collected in VTM. When appropriate and feasible, intestinal samples were also taken from deceased animals. Samples were aliquoted and stored at –80 °C until use. Blood samples were collected from recovered sows and workers on the farms who had close contact with sick pigs. Serum was separated by centrifugation at 3,000g for 15 min within 24 h of collection and preserved at 4 °C. Human serum collection was approved by the Medical Ethics Committee of the Wuhan School of Public Health, Wuhan University and Hummingbird IRB. Human, pigs and bats were sampled without gender or age preference unless indicated (for example, piglets or sows). No statistical methods were used to predetermine sample size.

Virus isolation. The following cells were used for virus isolation in this study: Vero (cultured in DMEM and 10% FBS); *Rhinolophus sinicus* primary or immortalized cells generated in our laboratory (all cultured in DMEM/F12 and 15% FBS): kidney primary cells (RsKi9409), lung primary cells (RsLu4323), lung immortalized cells (RsLuT), brain immortalized cells (RsBrT) and heart immortalized cells (RsHeT); and swine cell lines: two intestinal porcine enterocytes cell lines, IPEC (RPMI1640 and 10% FBS) and SIEC (DMEM and 10% FBS), three kidney cell lines PK15, LLC-PK1 (DMEM and 10% FBS for both) and IBRS (MEM and 10% FBS), and one pig testes cell line, ST (DMEM and 10% FBS). All cell lines were tested free of mycoplasma contamination, species were confirmed and authenticated by microscopic morphologic evaluation. None of the cell lines was on the list of commonly misidentified cell lines (by the ICLAC).

Cultured cell monolayers were maintained in their respective medium. PCR-positive pig faecal samples or the supernatant from homogenized pig intestine (in 200 μl VTM) were spun at 8,000g for 15 min, filtered and diluted 1:2 with DMEM supplemented with 16 μg ml^{-1} trypsin before addition to the cells. After incubation at 37 °C for 1 h, the inoculum was removed and replaced with fresh culture medium containing antibiotics (below) and 16 μg ml^{-1} trypsin. The cells were incubated at 37 °C and observed daily for cytopathic effect (CPE). Four blind passages (three-day interval between every passage) were performed for each sample. After each passage, both the culture supernatant and cell pellet were examined for the presence of virus by RT–PCR using the SADS-CoV primers listed in Supplementary Table 2. Penicillin (100 units ml^{-1}) and streptomycin (15 μg ml^{-1}) were included in all tissue culture media.

RNA extraction, *S1* gene amplification and qPCR. Whenever commercial kits were used, the manufacturer's instructions were followed without modification. RNA was extracted from 200 μl of swab samples (bat), faeces or homogenized intestine (pig) with the High Pure Viral RNA Kit (Roche). RNA was eluted in 50 μl of elution buffer and used as the template for RT–PCR. Reverse transcription was performed using the SuperScript III kit (Thermo Fisher Scientific).

To amplify *S1* genes from bat samples, nested PCR was performed with primers designed based on HKU2-CoV (GenBank accession number NC_009988.1)[19] (Supplementary Table 2). The 25-μl first-round PCR mixture contained 2.5 μl 10× PCR reaction buffer, 5 pmol of each primer, 50 mM MgCl$_2$, 0.5 mM dNTP, 0.1 μl Platinum Taq Enzyme (Thermo Fisher Scientific) and 1 μl cDNA. The 50-μl second-round PCR mixture was identical to the first-round PCR mixture except for the primers. Amplification of both rounds was performed as follows: 94 °C for 5 min followed by 60 cycles at 94 °C for 30 s, 50 °C for 40 s, 72 °C for 2.5 min, and a final extension at 72 °C for 10 min. PCR products were gel-purified and sequenced.

For qPCR analysis, primers based on SADS-CoV *RdRp* and *N* genes were used (Supplementary Table 2). RNA extracted from above was reverse-transcribed using PrimeScript RT Master Mix (Takara). The 10 μl qPCR reaction mix contained 5 μl 2× SYBR premix Ex TaqII (Takara), 0.4 μM of each primer and 1 μl cDNA. Amplification was performed as follows: 95 °C for 30 s followed by 40 cycles at 95 °C for 5 s, 60 °C for 30 s, and a melting curve step.

Luciferase immunoprecipitation system assay. The SADS-CoV *S1* gene was codon-optimized for eukaryotic expression, synthesized (GenScript) and cloned in frame with the Renilla luciferase gene (Rluc) and a Flag tag in the pREN2 vector[21]. pREN2-S1 plasmids were transfected into Cos-1 cells using Lipofectamine 2000 (Thermo Fisher Scientific). At 48 h post-transfection, cells were collected, lysed and a luciferase assay was performed to determine Rluc expression for both the empty vector (pREN2) and the pREN2-S1 construct. For testing of unknown pig or human serum samples, 1 μl of serum was incubated with 10 million units of Rluc alone (vector) or Rluc-S1, respectively, together with 3.5 μl of a 30% protein A/G UltraLink resin suspension (Pierce, Thermo Fisher Scientific). After extensive washing to remove unbounded luciferase-tagged antigens, the captured luciferase amount was determined using the commercial luciferase substrate kit (Promega). The ratio of Rluc-S1:Rluc (vector) was used to determine the specific S1 reactivity of pig and human sera. Commercial Flag antibody (Thermo Fisher Scientific)

was used as the positive control, and various pig sera (from uninfected animals in China or Singapore; or pigs infected with PEDV, TGEV or Nipah virus) were used as a negative control.

Protein expression and antibody production. The *N* gene from SADSr-CoV 3755 (GenBank accession number MF094702), which shares a 98% amino acid sequence identity to the SADS-CoV N protein, was inserted into pET-28a+ (Novagen) for prokaryotic expression. Transformed *Escherichia coli* were grown at 37 °C for 12–18 h in medium containing 1 mM IPTG. Bacteria were collected by centrifugation and resuspended in 30 ml of 5 mM imidazole and lysed by sonication. The lysate, from which N protein expression was confirmed with an anti-His-tag antibody, was applied to Ni^{2+} resin (Thermo Fisher Scientific). The purified N protein, at a concentration of 400 μg ml^{-1}, was used to immunize rabbits for antibody production following published methods[27]. After immunization and two boosts, rabbits were euthanized and sera were collected. Rabbit anti-N protein serum was used 1:10,000 for subsequent western blots.

Amplification, cloning and expression of human and swine genes. Construction of expression clones for human *ACE2* in pcDNA3.1 has been described previously[5,28]. Human *DPP4* was amplified from human cell lines. Human *APN* (also known as *ANPEP*) was commercially synthesized. Swine *APN* (also known as *ANPEP*), *DPP4* and *ACE2* were amplified from piglet intestine. Full-length gene fragments were amplified using specific primers (provided upon request). Human *ACE2* was cloned into pCDNA3.1 fused with a His tag. Human *APN* and *DPP4*, swine *APN*, *DPP4* and *ACE2* were cloned into pCAGGS fused with an S tag. Purified plasmids were transfected into HeLa cells. After 24 h, expression human or swine genes in HeLa cells was confirmed by immunofluorescence assay using mouse anti-His tag or mouse anti-S tag monoclonal antibodies (produced in house) followed by Cy3-labelled goat anti-mouse/rabbit IgG (Proteintech Group).

Pseudovirus preparation. The codon-humanized *S* genes of SADS-CoV or MERS-CoV cloned into pcDNA3.1 were used for pseudovirus construction as described previously[5,28]. In brief, 15 μg of each pHIV-Luc plasmid (pNL4.3.Luc.R-E-Luc) and the S-protein-expressing plasmid (or empty vector control) were co-transfected into 4×10^6 HEK293T cells using Lipofectamine 3000 (Thermo Fisher Scientific). After 4 h, the medium was replaced with fresh medium. Supernatants were collected 48 h after transfection and clarified by centrifugation at 3,000g, then passed through a 0.45-μm filter (Millipore). The filtered supernatants were stored at −80 °C in aliquots until use. To evaluate the incorporation of S proteins into the core of HIV virions, pseudoviruses in supernatant (20 ml) were concentrated by ultracentrifugation through a 20% sucrose cushion (5 ml) at 80,000g for 90 min using a SW41 rotor (Beckman). Pelleted pseudoviruses were dissolved in 50 μl phosphate-buffered saline (PBS) and examined by electron microscopy.

Pseudovirus infection. HeLa cells transiently expressing APN, ACE2 or DPP4 were prepared using Lipofectamine 2000 (Thermo Fisher Scientific). Pseudoviruses prepared above were added to HeLa cells overexpressing APN, ACE2 or DPP4 24 h after transfection. The unabsorbed viruses were removed and replaced with fresh medium at 3 h after infection. The infection was monitored by measuring the luciferase activity conferred by the reporter gene carried by the pseudovirus, using the Luciferase Assay System (Promega) as follows: cells were lysed 48 h after infection, and 20 μl of the lysates was taken for determining luciferase activity after the addition of 50 μl of luciferase substrate.

Examination of known CoV receptors for SADS-CoV entry/infection. HeLa cells transiently expressing APN, ACE2 or DPP4 were prepared using Lipofectamine 2000 (Thermo Fisher Scientific) in a 96-well plate, with mock-transfected cells as controls. SADS-CoV grown in Vero cells was used to infect HeLa cells transiently expressing APN, ACE2 or DPP4. The inoculum was removed after 1 h of absorption and washed twice with PBS and supplemented with medium. SARS-related-CoV WIV16[7] and MERS-CoV HIV-pseudovirus were used as positive control for human/swine ACE2 or human/swine DPP4, respectively. After 24 h of infection, cells were washed with PBS and fixed with 4% formaldehyde in PBS (pH 7.4) for 20 min at room temperature. SARS-related-CoV WIV16 replication was detected using rabbit antibody against the SARS-related-CoV Rp3 N protein (made in house, 1:100) followed by Cy3-conjugated goat anti-rabbit IgG (1:50, Proteintech)[7]. SADS-CoV replication was monitored using rabbit antibody against the SADSr-CoV 3755 N protein (made in house, 1:50) followed by FITC-conjugated goat anti-rabbit IgG (1:50, Proteintech). Nuclei were stained with DAPI (Beyotime). Staining patterns were examined using confocal microscopy on a FV1200 microscope (Olympus). Infection of MERS-CoV HIV-pseudovirus was monitored by luciferase 48 h after infection.

High-throughput sequencing, pathogen screening and genome assembly. Tissue from the small intestine of deceased pigs was homogenized and filtered through 0.45-μm filters before nucleic acid extraction and ribosomal RNA was depleted using the NEBNext rRNA Depletion Kit (New England Biolabs). Metagenomics analysis of both RNA and DNA viruses was performed. For RNA virus screening, the sequencing library was constructed using Ion Total RNA-Seq Kit v2 (Thermo Fisher Scientific). For DNA virus screening, NEBNext Fast DNA Fragmentation

& Library Prep Set for Ion Torrent (New England Biolabs) was used for library preparation. Both libraries were sequenced on an Ion S5 sequencer (Thermo Fisher Scientific). An analysis pipeline was applied to the sequencing data, which included the following analysis steps: (1) raw data quality filtering; (2) host genomic sequence filtering; (3) BLASTn search against the virus nucleotide database using BLAST; (4) BLASTx search against the virus protein database using DIAMOND v.0.9.0; (5) contig assembling and BLASTx search against the virus protein database. For whole viral genome sequencing, amplicon primers (provided upon request) were designed using the Thermo Fisher Scientific online tool with the HKU2-CoV and the SADS-CoV farm A genomes as references, and the sequencing libraries were constructed using NEBNext Ultra II DNA Library Prep Kit for Illumina and sequenced on an MiSeq sequencer. PCR and Sanger sequencing was performed to fill gaps in the genome. Genome sequences were assembled using CLC Genomic Workbench v.9.0. 5′-RACE was performed to determine the 5′-end of the genomes using SMARTer RACE 5′/3′ Kit (Takara). Genomes were annotated using Clone Manager Professional Suite 8 (Sci-Ed Software).

Phylogenetic analysis. SADS-CoV genome sequences and other representative coronavirus sequences (obtained from GenBank) were aligned using MAFFT v.7.221. Phylogenetic analyses with full-length genome, *S* gene and *RdRp* were performed using MrBayes v.3.2. Markov chain Monte Carlo was run for 20–50 million steps using the GTR+G+I model (general time reversible model of nucleotide substitution with a proportion of invariant sites and γ-distributed rates among sites). The first 10% was removed as burn-in. The association between phylogenies and phenotypes (for example, host species and farms) was assessed by BaTS beta-build2, with the trees obtained in the previous step used as input. For SADS-CoVs, a median-joining network analysis was performed using PopART v.1.7, with $\varepsilon = 0$. Phylogenetic analysis of the 33 full-length SADS-CoV genome sequences was performed using RAxML v.8.2.11, with GTRGAMMA as the nucleotide substitution model and 1,000 bootstrap replicates. The maximum likelihood tree was used to test the molecular clock using TempEst v.1.5. Potential genetic recombination events in our datasets were detected using RDP v.4.72.

Animal infection studies. Experiments were carried out strictly in accordance with the recommendations of the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The use of animals in this study was approved by the South China Agricultural University Committee of Animal Experiments (approval number 201004152).

Two different animal challenge experiments were conducted. Pigs were used without gender preference. In the first experiment, which was conducted before the virus was isolated, we used three-day old specific pathogen-free (SPF) piglets of the same breeding line, cared for at a SPF facility, fed with colostrum (except one). These piglets were bred and reared to be free of PEDV, CSFV, SIV, PCV2 and PPV infections, and were routinely tested for viral infections using PCR. We also conducted NGS to further confirm that these were animals were free of infection of the above viruses before the animal experiment, and to demonstrate that the animals were free of SADS-CoV infection. The intestinal tissue samples from healthy and diseased animals (intestinal samples excised from euthanized piglets, then ground to make slurry for the inoculum and NGS was performed to confirm no other pig pathogens were found in the samples), were used to feed two groups of 5 (control) and 7 (infection) animals, respectively. For the second experiment, isolated SADS-CoV was used to infect healthy piglets from a farm in Guangdong, which had been free of diarrheal disease for a number of weeks. These piglets were

from the same breed as those on SADS-affected farms, to eliminate potential host factor differences and to more accurately reproduce the conditions that occurred during the outbreak in the region. Both groups of piglets were cared for at a known pig disease-free facility. Again, qPCR and NGS were used to make sure that there was no other known swine diarrhoea virus present in the virus inoculum or any of the experimental animals. Two groups (6 for each group) of three-day old piglets were inoculated with SADS-CoV culture supernatant or normal cell culture medium as control. NGS and qPCR were used to confirm that there were no other known swine pathogens in the inoculum.

For both experiments, animals were recorded daily for signs of diseases, such as diarrhoea, weight loss and death. Faecal swabs were collected daily from all animals and screened for known swine diarrhoea viruses by qPCR. Weight loss was calculated as the percentage weight loss compared the original weight at day 0 with a threshold of >5%. It is important to point out that piglets when they are three days old tend to suffer from diarrhoea and weight loss when they are taken away from sows and the natural breast-feeding environment even without infection. At experimental endpoints, piglets were humanely euthanized and necropsies performed. Pictures were taken to record gross pathological changes to the intestines. Ileal, jejunal and duodenal tissues were taken from selected animals and stored at −80 °C for further analysis.

Haematoxylin and eosin and immunohistochemistry analysis. Frozen (−80 °C) small intestinal tissues including duodenum, jejunum and ileum taken from the experimentally infected pigs were pre-frozen at −20 °C for 10 min. Tissues were then embedded in optimal cutting temperature (OCT) compound and cut into 8-μm sections using the Cryotome FSE machine (Thermo Fisher Scientific). Mounted microscope slides were fixed with paraformaldehyde and stained with haematoxylin and eosin for histopathological examination.

For immunohistochemistry analysis, a rabbit antibody raised against the SADSr-CoV 3755 N protein was used for specific staining of SADS-CoV antigen. Slides were blocked by incubating with 10% goat serum (Beyotime) at 37 °C for 30 min, followed by overnight incubation at 4 °C with the rabbit anti-3755 N protein serum (1:1,000) and mouse anti-cytokeratin 8+18+19 monoclonal antibody (Abcam), diluted 1:100 in PBST buffer containing 5% goat serum. After washing, slides were then incubated for 50 min at room temperature with Cy3-conjugated goat-anti-rabbit IgG (Proteintech) and FITC-conjugated goat-anti-mouse IgG (Proteintech), diluted 1:100 in PBST buffer containing 5% goat serum. Slides were stained with DAPI (Beyotime) and observed under a fluorescence microscope (Nikon).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequence data that support the findings of this study have been deposited in GenBank with accession codes MF094681–MF094688, MF769416–MF769444, MF094697–MF094701, MF769406–MF769415 and MG557844. Raw sequencing data that support the findings of this study have been deposited in the Sequence Read Achieve (SRA) with accession codes SRR5991648, SRR5991649, SRR5991650, SRR5991651, SRR5991652, SRR5991654, SRR5991655, SRR5991656, SRR5991657, SRR5991658 and SRR5995595.

27. Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1988).
28. Ren, W. et al. Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin. *J. Virol.* **82**, 1899–1907 (2008).
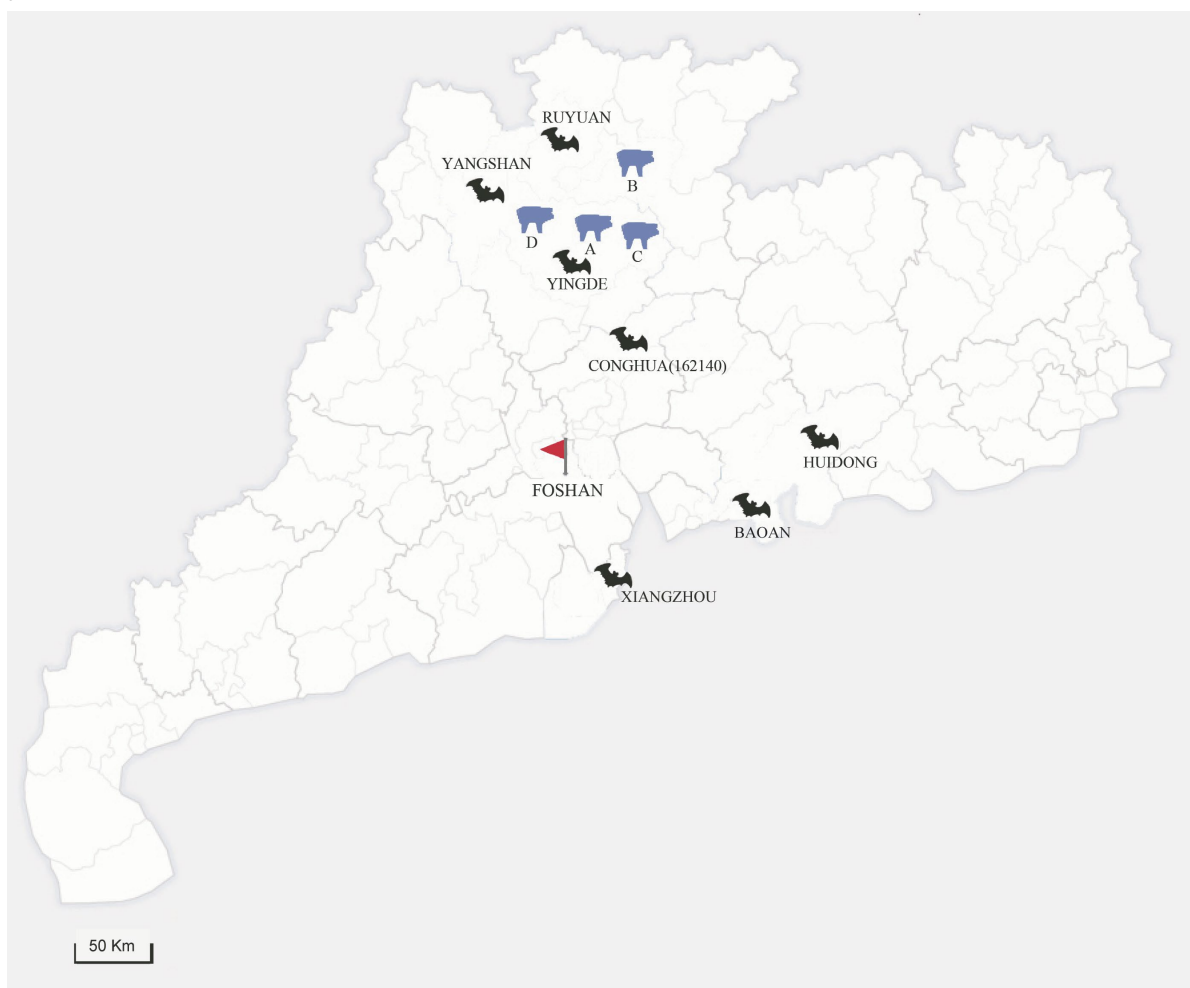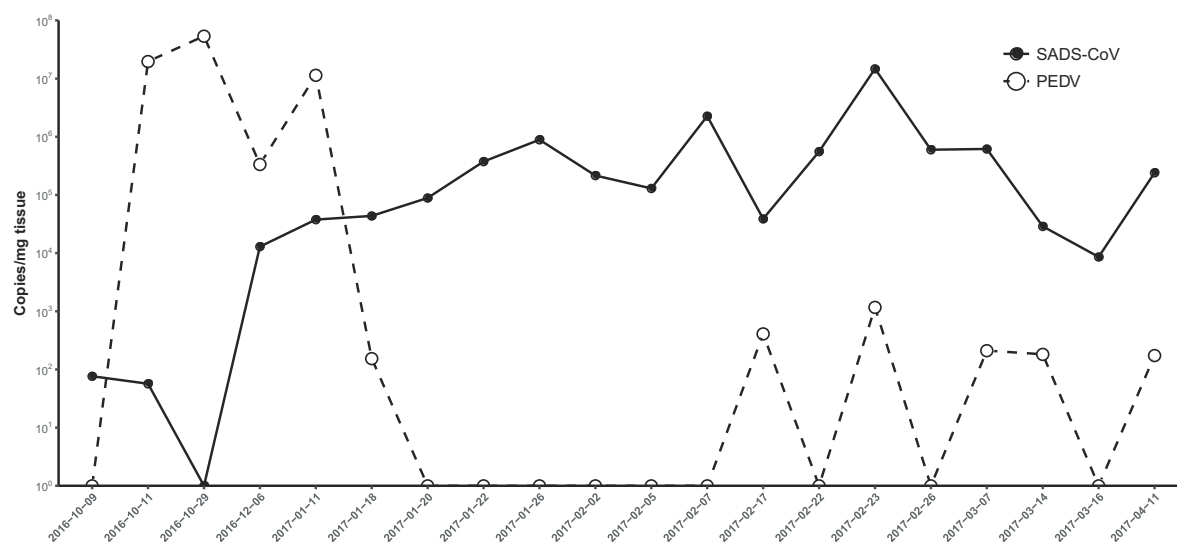
Extended Data Fig. 1 | Map of outbreak locations and sampling sites in Guangdong province, China and the co-circulation of PEDV and SADS-CoV during the initial outbreak on farm A. a, SADS-affected farms are labelled (farms A–D) with blue swine silhouettes following the temporal sequence of the outbreaks. Bat sampling sites are indicated with black bat silhouettes. The bat SADSr-CoV that is most closely related to SADS-CoV (sample 162140) originated in Conghua. The red flag marks Foshan city, the site of the SARS index case. **b**, Pooled intestinal samples (*n* = 5 or more biological independent samples) were collected at dates given on the *x* axis from deceased piglets and analysed by qPCR. The viral load for each piglet is shown as copy number per milligram of intestine tissue (*y* axis).
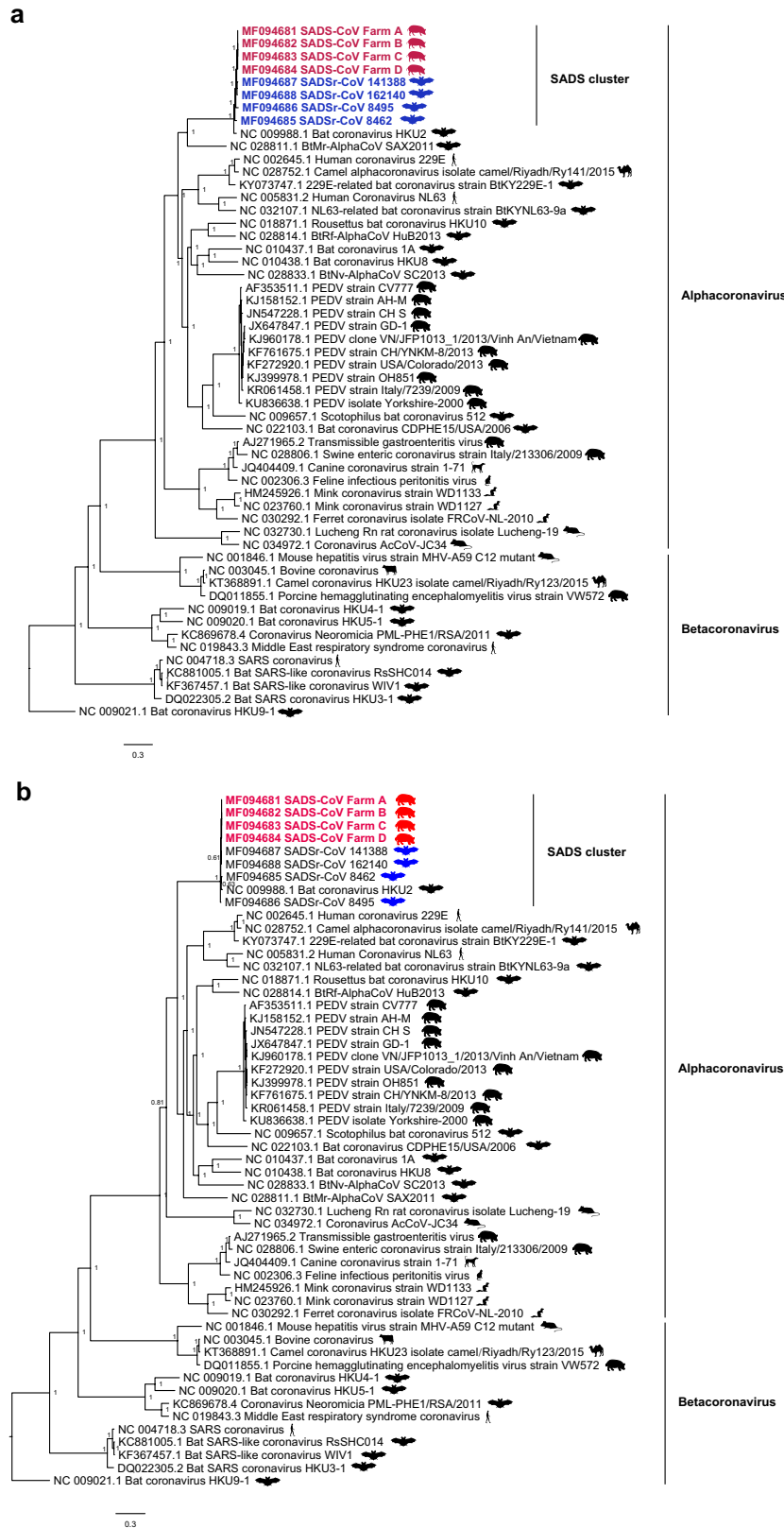
Extended Data Fig. 2 | Bayesian phylogenetic tree of the full-length genome and the *ORF1a* and *ORF1b* sequences of SADS-CoV and related coronaviruses. a, Bayesian phylogenetic tree of the full-length genome. **b**, Bayesian phylogenetic tree of the *ORF1a* and *ORF1b* sequences. Trees were constructed using MrBayes with the average standard deviation of split frequencies under 0.01. The host of each sequence is represented as a silhouette. Newly sequenced SADS-CoVs are highlighted in red, bat SADSr-CoVs are shown in blue and previously published sequences are shown in black. Scale bars, nucleotide substitutions per site.
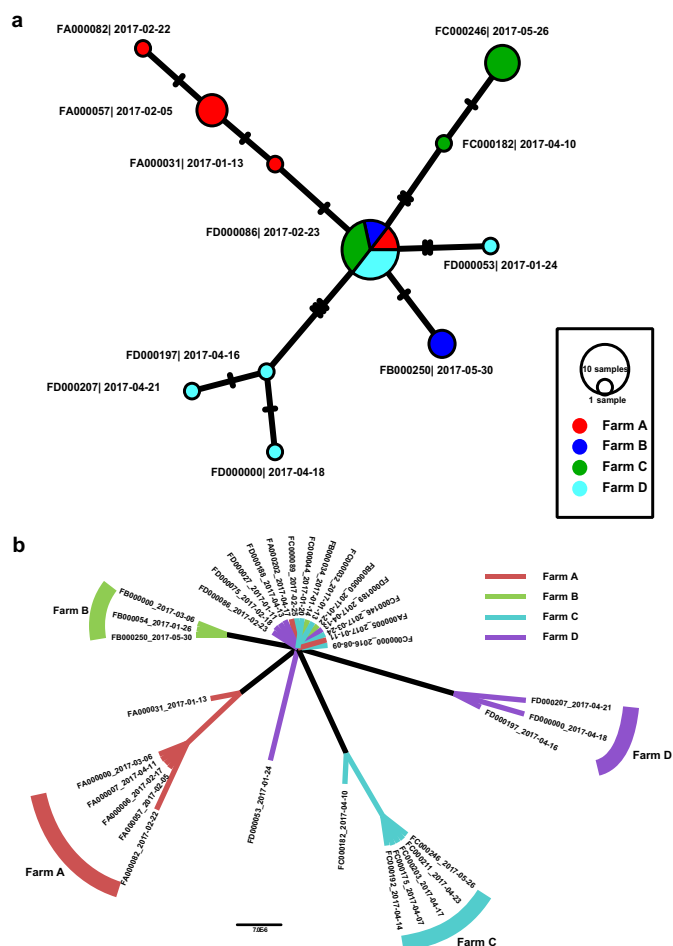
Extended Data Fig. 3 | Phylogeny and haplotype network analyses of the 33 SADS-CoV strains from the four farms. a, Phylogenetic tree constructed using MrBayes. The GTR+GAMMA model was applied and 20 million steps were run, with the first 10% removed as burn in. Viruses from different farms are labelled with different colours. Scale bar, nucleotide substitutions per site. **b**, Median-joining haplotype network constructed using ProART. In this analysis, $\varepsilon = 0$ was used. The size of the circles represents the number of samples. The larger the circle, the more samples it includes.

Extended Data Fig. 4 | Recombination analysis for SADS-CoV and related CoVs. The potential genetic recombination events were detected using RDP. For each virus strain, different colours represent different sources of the genomes.
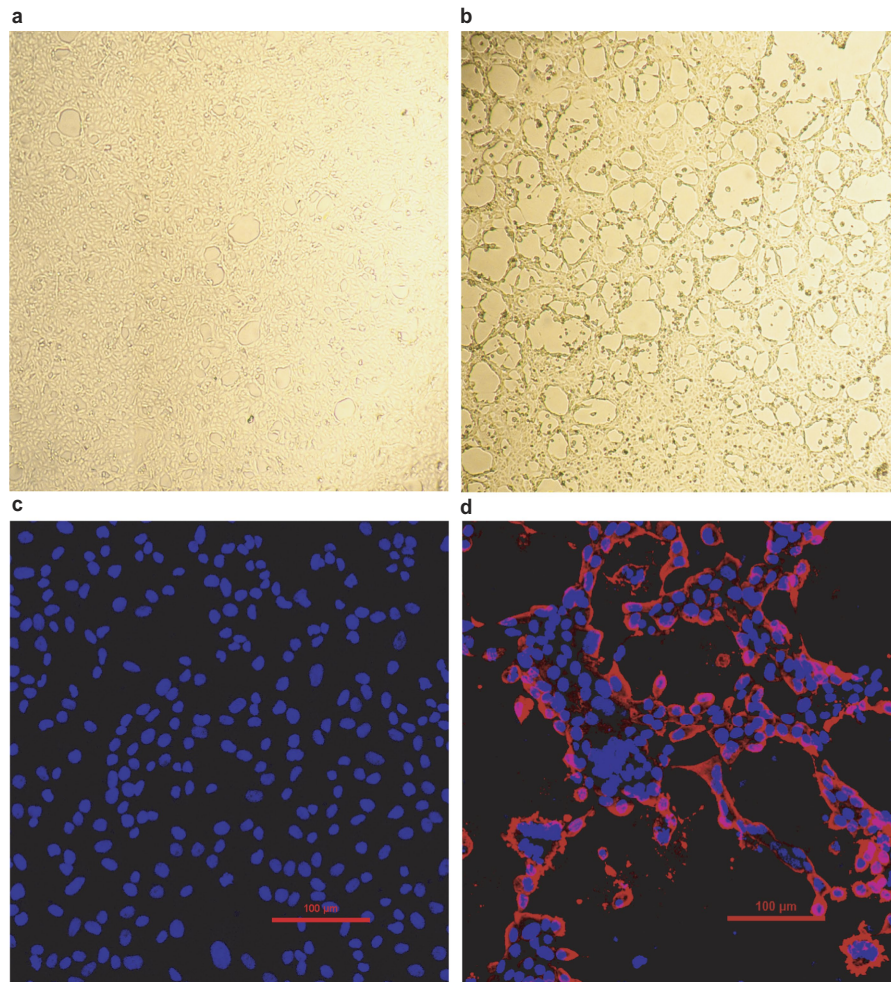
Extended Data Fig. 5 | Isolation and antigenic characterization of SADS-CoV. a, b, Vero cells are shown 20 h after infection with mock (**a**) or SADS-CoV (**b**). **c, d,** Mock or SADS-CoV-infected samples stained with rabbit serum raised against the recombinant SADSr-CoV N protein (red) and DAPI (blue). The experiment was conducted independently three times with similar results. Scale bars, 100 μm.

Extended Data Table 1 | List of all known swine viruses tested by PCR at the beginning of the of SADS outbreak investigation on the four farms

	PEDV	PDCoV	TGEV	RV	PBV	PSV	SVA	SIV	NADC30	PRV	FMDV	CSFV	PCV2	PCV3	APPV	PPV	Norovirus
Farm A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	ND	-
Farm B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	ND	-
Farm C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	ND
Farm D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	ND

Faeces, intestine or faecal swabs collected from January to April 2017 were tested. Sampling type and number of samples per farm were as follows. Farm A: 1 fecal sample, 20 intestinal sample and 6 faecal swabs; farm B: 1 faecal sample and 15 intestinal samples; farm C: 2 intestinal sample and 1 faecal swab; farm D: 5 faecal sample and 1 faecal swab. The dash indicates a negative PCR result. ND, not determined. APPV, atypical porcine pestivirus; CSFV, classical swine fever virus; FMDV, foot and mouth disease virus; NADC30, porcine reproductive and respiratory syndrome virus, strain NADC30; PBV, porcine picobirnavirus; PCV2, porcine circovirus 2; PCV3, porcine circovirus 3; PDCoV, porcine deltacoronavirus; PEDV, porcine epidemic diarrhoea virus; PPV, porcine parvovirus; PRV, porcine pseudorabies virus; PSV, porcine sapelovirus; RV, porcine rotavirus; SIV, swine influenza virus; SVA, porcine senecavirus A; TGEV, porcine transmissible gastroenteritis virus.

Extended Data Table 2 | List of SADSr-CoVs detected in bats in Guangdong, China

Sampling			PCR analysis	
Time (Month-Year)	Location	Bat Species	Fecal swabs sampled	PCR Positive
Jun 13	Yingde	*Rhinolophus sinicus*	1	0
		Pipistrellus abramus	8	0
		Myotis ricketti	2	0
Jul 13	Yangshan	*Pipistrellus abramus*	1	0
		Hipposideros pratti	36	0
Jul 13; May 14; Jun 15; Aug 16	Ruyuan	*Rhinolophus sinicus*	27	5
		Rhinolophus affinis	11	0
		Rhinolophus macrotis	3	0
		Rhinolophus pusillus	41	3
		Rhinolophus rex	9	7
		Hipposideros pratti	7	0
Sep 14; Jun 15; Aug 16	Conghua	*Rhinolophus sinicus*	70	2
		Rhinolophus affinis	34	7
		Rhinolophus pusillus	11	2
		Hipposideros pomona	10	0
		Myotis ricketti	1	0
Jun 13; Nov 13; Aug 14; Jun 15	Huidong	*Rhinolophus sinicus*	37	4
		Rhinolophus affinis	59	27
		Rhinolophus macrotis	15	0
		Rhinolophus pusillus	1	0
		Hipposideros pomona	2	0
		Myotis ricketti	84	0
Jun 15	Baoan	*Rhinolophus sinicus*	55	1
Sep 14	Xiangzhou	*Rhinolophus pusillus*	28	0
		Hipposideros pomona	38	0
		Total	591	58 (9.8%)

See Extended Data Fig. 1 for sampling sites in relation to SARS and SADS outbreak locations.

Extended Data Table 3 | Test of SADS-CoV entry and infection in Hela cells expressing known coronavirus receptors

	HuAPN*	HuACE2*	HuDPP4*	SwAPN*	SwACE2*	SwDPP4*
SADS-CoV	-	-	-	-	-	-
SARS-related-CoV	NA	+	NA	NA	+	NA
MERS-CoV[†]	NA	NA	+	NA	NA	NA
Expression[‡]	+ (S-tag)	+ (HIS-tag)	+ (S-tag)	+ (S-tag)	+ (S-tag)	+ (S-tag)

*Gene accession numbers for the genes used in this study: human APN, M22324.1; human ACE2, NM_021804; human DPP4, NM_001935.3; SwAPN (swine APN), NM_214277.1; SwACE2 (swine ACE2), NM_001116542.1; SwDPP4 (swine DPP4), NM_214257.1.
†For MERS-CoV infection, HIV-pseudovirus was used.
‡Expression of APN, ACE2 and DPP4 was confirmed by antibodies against fused tags.

Extended Data Table 4 | Experimental infection of SPF piglets using intestine tissue homogenate

a

Group	Animal Number	Age (days)	Inoculum material	SADS-CoV titer (copy/ml)	Inoculum volume	Inoculation route	Data recorded on day one and (day two) post challenge				
							Death	Weight loss	Watery diarrhea	SADS-CoV (+ve)	PEDV/PDCoV/RV (+ve)
Infected	7	3	PCR positive intestine slurry	1.55×10*6	4 ml	Oral + milk	0/7 (3/7)	4/7 (5/7)	5/7 (7/7)	6/7 (7/7)	0/7 (0/7)
Control	5	3	PCR negative intestine slurry	0	4 ml	Oral + milk	0/5 (1/5)	1/5 (3/5)	0/5 (1/5)	0/5 (0/5)	0/5 (0/5)

b

Group	Days post challenge	Piglet-I1*	Piglet-I2*	Piglet-I3*	Piglet-I4*	Piglet-I5†	Piglet-I6†	Piglet-I7†
Infected	0	0.565	0.66	0.6	0.68	0.49	0.57	0.62
	1	0.555	0.635	0.685	0.715	0.4	0.475	0.565
	2	0.51	0.52	0.665	0.785			

Group	Days post challenge	Piglet-C1*	Piglet-C2*	Piglet-C3*	Piglet-C4†‡	Piglet-C5*
Control	0	0.67	0.59	0.5	0.53	0.525
	1	0.765	0.53	0.49	0.51	0.535
	2	0.765	0.53	0.575		0.505

Experimental details can be found in the Methods. **a**, Animals were recorded every day for signs of disease, including weight loss, diarrhoea and death. PCR on DNA from faecal swabs was carried out to monitor the presence of SADS-CoV or other pig viruses. **b**, Daily body weight record of all piglets. Weights are in kg.
*Euthanized on the indicated day for further analysis.
†Animal died during the experiment.
‡The only animal that did not receive colostrum in this experiment due to shortage in supply.

Extended Data Table 5 | Experimental animal infection of farm piglets using cultured SADS-CoV

a

Group	Animal Number	Age (days)	Inoculum material	SADS-CoV titer (TCID$_{50}$/ml)	Inoculum volume	Inoculation route	Data recorded on day two and (day four) post challenge				
							Death	Weight loss	Watery diarrhea	SADS-CoV (+ve)	PEDV/PDCoV/RV (+ve)
Infected	6	3	Cultured SADS-CoV	10*6.625	6 ml	Oral + milk	1/6 (3/6)	4/6 (6/6)	6/6 (6/6)	6/6 (6/6)	0/6 (0/6)
Control	6	3	Mock culture supernatant	0	6 ml	Oral + milk	0/6 (0/6)	3/6 (3/6)	5/6 (3/6)	0/6 (0/6)	0/6 (0/6)

b

Group	Days post challenge	Piglet-I1[†]	Piglet-I2[†]	Piglet-I3*	Piglet-I4*	Piglet-I5*	Piglet-I6[†]
Infected	0	1.5	1.54	2.32	1.92	1.54	2.165
	1	1.41	1.575	2.58	1.885	1.46	2.08
	2	1.23	1.39	2.615	1.73	1.54	1.365
	3			2.115	1.54	1.335	1.725
	4						1.505

Group	Days post challenge	Piglet-C1*	Piglet-C2*	Piglet-C3*	Piglet-C4*	Piglet-C5*	Piglet-C6*
Control	0	1.955	2.055	2.8	1.835	1.835	1.83
	1	1.765	1.955	1.9	1.68	1.645	1.93
	2		2.12	1.675	1.93	1.515	1.9
	3		2.25	1.69	2.18	1.66	2.38
	4				2.27	1.555	2.58

Experimental details can be found in the Methods. **a**, Animals were recorded every day for signs of disease, including weight loss, diarrhoea and death. PCR on DNA from faecal swabs was carried out to monitor the presence of SADS-CoV or other pig viruses. **b**, Daily body weight record of all piglets. Weights are in kg.
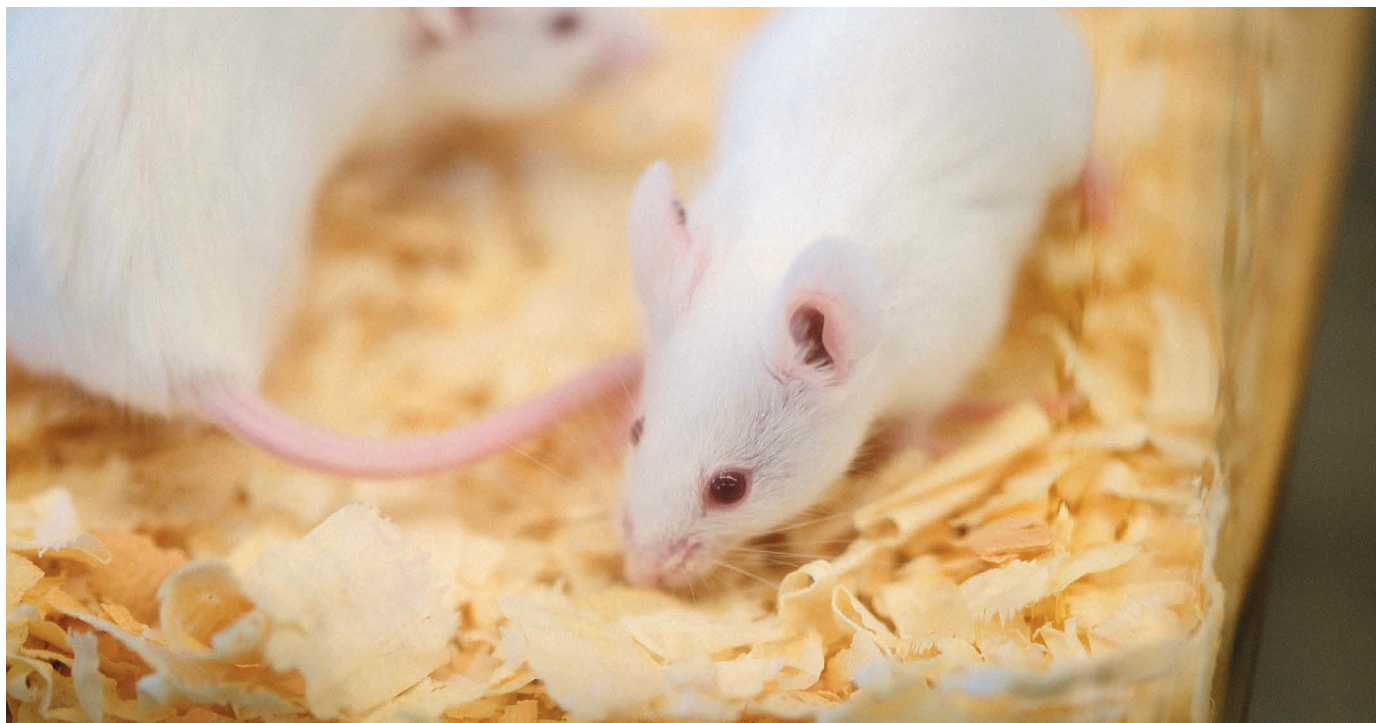*Euthanized on the indicated day for further analysis.
†Animal died during the experiment.

MOUSE MODELS WITH A HUMAN TOUCH

Engineered mice are valuable for disease and drug research, but scientists hunger for cancer models that better mirror the condition in humans.

WHITNEY HAYWARD/PORTLAND PRESS HERALD/GETTY

Mice are commonly used to study cancer, but scientists are still working to improve modelling of the human disease.

BY MIKE MAY

In 1915, with the world at war, Japanese pathologist Katsusaburo Yamagiwa and his assistant Koichi Ichikawa were focused on a killer nearly as deadly as the battle raging on the Western Front. The duo, based at what was then the Imperial University of Tokyo, had spent more than 150 days painting coal tar on the ears of rabbits. Finally, they found that the rabbits had cancer.

Yamagiwa's diseased rabbits are considered to have been the first animal model for cancer research[1]. Since then, scientists have used everything from cell lines to engineered mice to try to mimic human cancer. But finding the option best suited to answering a specific experimental question requires a lot of thought.

According to medical oncologist David Weinstock of the Dana–Farber Cancer Institute in Boston, Massachusetts, what makes

for a good cancer model is "a very complex question, and the simplest answer is it must be able to give me insight — truly answer the question that I want to ask. If it can't do that, I'm wasting my time."

For Nancy Boudreau, a branch chief at the US National Cancer Institute (NCI) in Bethesda, Maryland, a model's fidelity to the course of human cancer is key. "The more it recapitulates the human disease and progression, the better," she says. An ideal cancer model should replicate many of the features that occur in human cancer, including how it develops and progresses when facing a human immune system; how it metastasizes, or spreads from its primary source to other parts of the body; and how it reacts to therapy. That requires scientists to know the pros and cons of each cancer model, because none will answer every research question.

Some evidence suggests that, despite many options, no existing model of cancer is good

enough for developing therapeutics. According to a report co-authored by the international Biotechnology Innovation Organization that examined clinical trials from 2006 to 2015, cancer drugs fared the worst out of 15 disease groupings, progressing from phase I to approval only 5.1% of the time (see go.nature.com/2pxfn16). By contrast, success rates for haematology and infectious-disease therapeutics were 26.1% and 19.1%, respectively.

"If better preclinical models could improve the clinical translatability by just 10%, that would very much improve the quality of preclinical cancer research and translate into enormous savings for drug developers," says Hellmut Augustin, a specialist in vascular oncology at the German Cancer Research Center in Heidelberg.

Groups collaborate worldwide to improve these models. Scientists at the NCI, Cancer Research UK in London, the Wellcome Trust ▶

▶ Sanger Institute in Hinxton, UK, and the not-for-profit Hubrecht Organoid Technology in Utrecht, the Netherlands, for instance, have teamed up on an effort called the Human Cancer Models Initiative. It launched in 2016 with the goal of developing 1,000 new cancer models in cell lines for use by researchers around the world. Such projects suggest that many scientists agree on the value of expanding the pool of models.

MODIFIED MOUSE GENOMES

For many questions, the humble cultured cell provides sufficient insight. But these cells are typically grown in unnatural 2D formats that lack the conditions in which human cancers grow — especially, an immune system. This makes cultured cells ill-suited for modelling many aspects of disease. Instead, says cancer-systems biologist Shannon Hughes of the NCI, a good starting point for many investigations is a genetically engineered mouse (GEM). "They are well characterized and well controlled," she says.

For years, engineering a mouse required complicated processes to generate desired DNA, transform cells in culture and inject them into an embryo to modify its genes. But the options for making a GEM today, like most other genetic-modification applications, changed with the discovery of the CRISPR gene-editing system. "CRISPR has enabled more-subtle manipulations that were extremely challenging with previous technologies," says cancer biologist Lukas Dow of the Weill Cornell Medical College in New York City.

"For instance," Dow says, "it is now relatively straightforward to induce large chromosome rearrangements — inversions, deletions and translocations" — associated with disease. With CRISPR, scientists can even change a single base in a rodent's DNA. Base-by-base resolution offers "the ability to accurately recreate the precise mutations observed in human cancer", he notes. "Such detail has been largely ignored in model development thus far, but it is increasingly apparent that the devil is in the detail."

Taeyoung Koo, a genome engineer at South Korea's Institute for Basic Science, based in Daejeon, and her colleagues used CRISPR to target a mutation in non-small-cell lung cancer (NSCLC)[2]. They report that of human cases of NSCLC, 15% involve a change to just one DNA base — known as a single-nucleotide mutation — in the epidermal growth-factor receptor (*EGFR*) gene. Current treatment consists of drugs, such as gefitinib, that target the mutated protein produced by that gene.

Koo's team developed a CRISPR–Cas9 guide-RNA sequence that recognizes the most commonly mutated *EGFR* region, which accounts for more than 40% of *EGFR*-mutation-related NSCLC cases. They then implanted mice with human NSCLC tumours and targeted the mutation with CRISPR–Cas9 and a specific guide RNA. Their results showed that a properly designed guide RNA is sufficiently precise to

break the diseased sequence, yielding a potential therapeutic strategy. The "mutant allele-specific Cas9 can efficiently distinguish the *EGFR* mutant allele from the wild-type allele, leading to targeted oncogene disruption and cancer cell death", they report[2].

Although CRISPR shows remarkable target specificity, the result of its activity can be highly variable. So, if the goal is to create consistent and uniform genetic changes across all cases, nucleases such as Cas9 are a poor choice, says Dow. "The random nature of DNA repair in traditional CRISPR systems means that you have to deal with a significant amount of heterogeneity in cell populations."

GEMs have limitations, too, especially concerning the timing and heterogeneity of disease. "Mouse tumours progress incredibly fast," Hughes explains. That speed enables researchers to accelerate their experiments, but they fail to replicate the pace of disease in humans. Plus, she says, the tumours tend to be too homogeneous to reflect human disease properly: a GEM usually includes just one or two genetic changes, whereas human tumours often have many.

To address the lack of heterogeneity and produce a more human-like model, biomedical scientist Lorenzo Federico and his colleagues, working in the laboratory of systems biologist Gordon Mills at the University of Texas MD

Anderson Cancer Center in Houston, engineered a collection of transplantable grafts from primary breast tumors in transgenic mice[3]. The procedure yielded 12 new graft lines of mice — mouse models that can reliably produce specific types of cancer with a wide array of genetic changes. "Ideally, different primary tumours arising in different mice should be characterized by different molecular alterations to more closely reflect the genetics of human cancer," Federico says.

These models have already been used successfully as preclinical platforms for the assessment of targeted therapeutics, including inhibitors of molecular pathways involved in cancer. According to Federico, they are also well suited for studying the role of the immune system in tumorigenesis and therapeutics development. Yet, because these transplantable grafts were derived from engineered mouse tumours, he says, the results recorded from this approach "must be always taken with a grain of salt".

How cancer arises and progresses depends intimately on its interaction with the host immune system. Some of the most promising treatments, called immunotherapies, engineer a patient's immune system to attack a specific tumour. To study these therapies, scientists need mice with an intact immune system — better yet, a human one. That led to humanized mice.
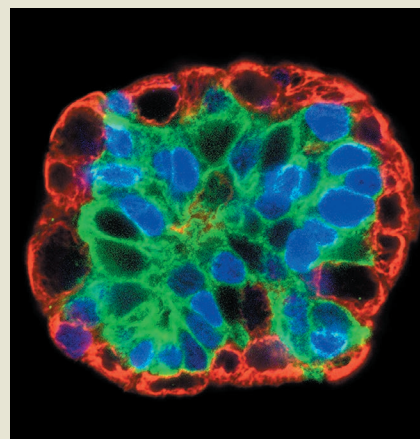
Organoid options

Organoids are an increasingly popular model.

Mice aren't the only options researchers have for modelling cancer. A popular emerging alternative is the organoid — a 3D cell culture that mimics some of the microanatomy of an organ, such as its system of blood vessels.

"A tumour is a kind of organ, where tissues cooperate," says molecular biologist Claudine Kieda of the Centre for Molecular Biophysics in Orleans, France. "A 3D cell model takes into account the microenvironment, such as the level of oxygen around and in the tumour."

Kieda's lab combines melanoma and endothelial cells in a matrix composed of collagen, growth factors and a 3D scaffold called Matrigel. This mixture allows the cells to form a structure that resembles a tumour and its surroundings, especially in terms of oxygenation[7]. "Everyone is working in conditions that are like an incubator, where the partial pressure of oxygen is much higher than in the body," Kieda says.

Among other uses, organoids are valuable for drug development. For instance, Meritxell Huch, a tissue-repair biologist at the University of Cambridge, UK, and her colleagues created liver-cancer organoids for drug screening[8]. This type of tumour can be grown in mice only about 20% of the time,

but Huch achieved a success rate of nearly 80% — and the process worked about twice as quickly as with a patient-derived xenograft. "Speed is the main advantage," Huch says. Using these organoids, Huch's team identified an inhibitor of a signalling pathway that represents a potential target for treating primary liver cancer.

As with other cancer models, a good organoid replicates the human disease as faithfully as possible. Organoids, says Nancy Boudreau, a metastasis researcher at the US National Cancer Institute in Bethesda, Maryland, "are more biological than cells in regular culture". And they are less expensive than mice. **M.M.**

For these GEMs, human hematopoietic stem cells — precursors to an array of blood cell types — are implanted into an immune-deficient mouse. This process recreates certain aspects of the human immune system, such as white blood cells called T cells, which attack foreign cells. Then, a sample of a human tumour — called a patient-derived xenograft (PDX) — can be transplanted into the mouse, creating a more realistic model of human disease.

According to Augustin, PDX models are increasingly popular among drug developers, who use them as test beds for drug testing. PDX models are also moving into basic-research labs, and are commercially available. Working with more than 20 cancer clinics, the Jackson Laboratory in Bar Harbor, Maine, has created more than 450 of these mouse models, including ones for acute myeloid leukaemia and bladder, breast, lung, ovarian and pancreatic cancer. They usually cost about three times as much as standard immune-deficient mice, the non-profit says.

Scientists can also develop their own PDX mice. Oncologist Elizabeth Stewart of St. Jude Children's Research Hospital in Memphis, Tennessee, and her colleagues used samples of surgically removed paediatric solid tumours, representing brain, bone and other cancers, to generate 67 PDX mouse models covering a dozen tumour types[4].

Stewart and her colleagues' aimed to create models for studying treatment efficacy against different tumour types — an approach that requires the model to represent the original disease accurately. Stewart's team decided to compare the PDX and source tumours at the nucleic-acid level using whole-genome and whole-exome DNA sequencing. Overall, they found, the PDX sequences largely matched the genomic features of the source tumours, although new mutations also emerged. The PDXs "retained the molecular and cellular features of the patient tumour and the epigenetic landscape of their developmental origins", the researchers concluded.

That's not to say that PDXs are static. Todd Golub, director of the cancer programme at the Broad Institute of Harvard and MIT in Cambridge, Massachusetts, and his team studied genomic rearrangements called copy-number variations (CNVs) in 543 PDX models representing 24 classes of cancer[5]. They found that expansive regions of CNVs constituting more than 5 million bases had been introduced into 60% of the PDXs after 1 passage from the original mouse to its offspring, and 88% of PDXs after 4 passages. The results show that PDXs that initially mimic human disease can evolve into forms that do not. When that happens, the PDX loses its faithfulness to the target cancer.

Boudreau describes engrafting human PDX-model tissues into humanized mice as one of the most intriguing new cancer models to emerge, but says it's "not quite there yet" because researchers have yet to learn the ins and outs of the technology. That said, she adds, the technology could prove useful for one hot facet of
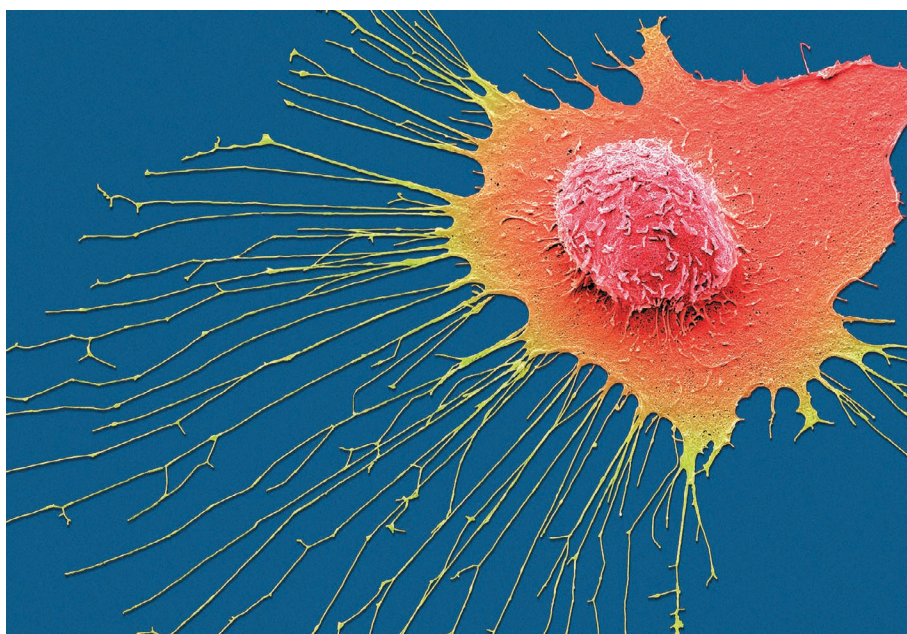
A breast-cancer cell migrates under the microscope.

therapeutics development: "Humanized mice will be pretty critical with the interest in immunotherapy and how human tumours respond," she says.

BACK TO THE BEGINNING

Rather than relying on genetic techniques to produce a better model of cancer, some scientists are going old school — using Yamagiwa's approach. This chemical-carcinogenesis method uses ordinary lab mice, and the results can create more-realistic cancer models.

"You treat a mouse with a carcinogen, like an environmental agent, to cause a specific type of damage and to get specific types of tumours, such as tumors in the skin," explains tumour biologist Melissa Reeves at the University of California, San Francisco. "This does a good job of recapitulating tumours in humans exposed to specific environments, because it models the natural evolution of a tumour caused by a wide array of genetic damage."

"The more it recapitulates the human disease and progression, the better."

Chemical carcinogens can damage DNA at hundreds of sites, and their impact can be followed over time. Reeves and her colleagues took this approach, using topical applications of known carcinogens called DMBA and TPA to induce skin cancer in mice, to study how tumours move from a primary site to a secondary one[6]. Her findings suggest that skin cancer does not travel serially from site to site — from skin to lymph nodes to lungs, for instance — but rather, "by parallel dissemination, going to the lymph nodes and lungs at the same time".

This finding, Reeves says, provides experimental validation of a well-documented clinical finding: that removing the lymph nodes around breast cancer doesn't always increase survival, an observation that led researchers to speculate about the possibility of parallel transmission.

Although chemical carcinogenesis creates diseases that, compared with GEMs and humanized mice, might better resemble the heterogeneity of human cancer, actually using these models has significant downsides. It can take 18 months to create a primary tumour through chemical means, remove it and study the course of metastasis. "Plus, every tumour is going to be a little bit different," Reeves says.

Likewise, every cancer model differs, and mice aren't always the best choice (see 'Organoid options'). Mice are expensive to maintain and pose ethical concerns, which will always make cell lines an option to consider.

For now, researchers must choose a model — despite its shortcomings — that they think will best answer their specific question. At the same time, scientists will keep advancing existing models and developing new ones. As Augustin notes, "It is well-invested money to develop and employ mouse tumour models with better translational relevance and impact." Otherwise, the performance of cancer drugs in clinical trials might never improve. ∎

Mike May *is a freelance writer based near Houston, Texas.*

1. Yamagiwa, K. & Ichikawa, K. J. *Cancer Res.* **27**, 123–181 (1918).
2. Koo, T., Yoon, A.-R., Cho, H.-Y., Bae, S., Yun, C.-O., & Kim, J.-S. *Nucleic Acids Res.* **45**, 7897–7908 (2017).
3. Federico, L. *et al. Sci. Adv.* **3**, e1600957 (2017).
4. Stewart, E. *et al. Nature* **549**, 96–100 (2017).
5. Ben-David, U. *et al. Nature Genet.* **49**, 1567–1575 (2017).
6. McCreery, M. *et al. Nature Med.* **21**, 1514–1520 (2015).
7. Klimkiewicz, K. *et al. Cancer Lett.* **396**, 10–20 (2017).
8. Broutier, L. *Nature Med.* **23**, 1424–1435 (2017).

CAREERS

MARK AIRS/GETTY

RESEARCH ENTERPRISE

The rise of outsourcing

Big pharma is downsizing, and contract research organizations are reaping the benefits.

BY ESTHER LANDHUIS

Alokta Chakrabarti manages drug-candidate identification and discovery for client pharmaceutical companies. As a project team leader at the contract research organization (CRO) ProQinase in Freiburg, Germany, she spends her days meeting clients,

working at the bench, flying through data analysis — and chasing a lot of deadlines.

A few decades ago, drug makers did their own discovery work, along with every other element of getting a drug or medical device to the marketplace. But today, nearly anything that a pharmaceutical, biotechnology or medical-device business needs to do — from designing

assays to planning and running clinical trials — can and may be outsourced to CROs.

These specialized companies fall into several categories. Preclinical CROs test drug or device candidates for client businesses before the compounds or devices undergo clinical or human testing. This might include helping a client to synthesize compounds, run biochemical assays or conduct animal studies. Clinical CROs focus on clinical-trial services, such as medical writing, data analysis, managing regulatory-affairs processes and other functions associated with getting new drugs or devices to market. A growing number of speciality CROs focus on a particular stage of clinical development, or offer services within a specific therapeutic niche.

The CRO industry is benefiting from recent downsizing trends in big pharma, as well as from a related proliferation of smaller drug makers, says Ken Getz, who studies research and development management practices at Tufts University in Boston, Massachusetts. The number of drugs entering clinical trials continues to rise, and companies that are slashing their workforce look to CROs to help them manage their portfolios. The need for outsourcing is even greater for smaller biopharmaceutical firms with lean headcounts and scant clinical experience, says Getz.

The global clinical CRO market topped US$23 billion in sales in 2014, and is predicted to exceed $35 billion in sales by 2020. More than one-third of all global drug-discovery research will be farmed out to CROs by 2021, predicts Kalorama Information, a market-research publisher in Rockville, Maryland, in its 2018 *Outsourcing in Drug Discovery* report (see go.nature.com/2jgjoqb).

As the CRO industry takes on larger and more-complex roles, the distinction between working in biopharma and at a CRO is blurring. Jobs in both areas are listed with life-sciences recruitment agencies and on job boards. Many people thought that few CROs could draw top researchers, says Josh Schultz, senior vice-president of Parexel, a full-service CRO headquartered near Boston, with a workforce of about 19,000 across more than 50 countries. Now, he says, those researchers are more evenly distributed. "CROs have become development partners in ways that we weren't before now," he says. "Client companies say, 'Can you take this compound from start to finish?'"

JOB BOOM

About 15 years ago, the top 5 CROs worldwide collectively employed around 30,000 people. Now, that group has nearly 100,000 ▶

employees, and a single CRO can have thousands of clinical trials in progress at any given time, estimates Schultz. According to the Association of Clinical Research Organizations (ACRO) in Washington DC, whose members run trials in 142 countries, more than half of CRO jobs are in the United States and Europe. India has 8% and the United Kingdom 7%, finds ACRO's 2015 member survey.

Researchers who have solid project-management and communication skills will be competitive for jobs at CROs — and could be even more strongly positioned if they have experience working with large data sets, say industry experts.

Chakrabarti joined ProQinase in 2016 after completing a postdoc in histone epigenetics at the University of Freiburg; before that, she had done undergraduate work in India and a cancer-biology PhD in the Netherlands. So it was a challenge to shift her mindset into a business-oriented, deadline-driven approach, she admits (see go.nature.com/2j0xipn). "You do your experiment, and if it does not work, you try another," she says of her PhD programme and postdoc. "There is no deadline."

Her CRO clients, however, often want results in one or two weeks. "You have to plan very wisely," Chakrabarti says. "If there's a problem, you have to troubleshoot. If you have too many projects, you end up with overlapping deadlines. It's a little stressful."

Many academic labs don't emphasize these kinds of project-management skills. "Particularly in PhD programmes, there's this culture saying you should learn how to do everything yourself," says Elizabeth Iorns, chief executive of Science Exchange in Palo Alto, California, a network of CROs, core labs and other scientific-service providers that runs experiments for a fee. But in industry, she says, you need to highlight and burnish specific skills and talents. "It's impossible to be trained in every technique," says Iorns.

Leaving academia requires a mental shift, Iorns says — less focus on creativity, invention and first-author papers, and more emphasis on discipline and quality control. Metrics for success also differ from those in an academic lab. "If your end goal is to bring a drug to market, you want to know as soon as possible that it's not going to work," she says. "Negative data are just as valuable as positive data."

CROs can offer excellent opportunities for those pursuing an undergraduate degree, say some, and they can provide a crash course in lab techniques and interpersonal skills. That was the case for Nikita Patel, who started a job at a local preclinical CRO after graduating from the University of San Diego in California. "CROs are often fast-paced, cut-throat environments that teach you a lot in a short time," Patel says. "Working there taught me to persevere and showed me the nitty-gritty nuances of doing science."

CROs can also be a good option for researchers who love science but hate benchwork. A few months into his master's degree in biology at Sonoma State University in California, Brian Wenzel realized that he wasn't suited for academia. "Lab work for me was really daunting," he says. "If you screw up halfway through, you could compromise two years' worth of data. It requires a certain kind of personality to be excited and tenacious enough to keep doing the same thing over and over with the precision needed for top-tier research."

So after graduating, in 2008, Wenzel started as a customer-service representative at a contract manufacturing organization that specialized in protein research, and moved on to other CRO sales and business positions. Patel, too, pivoted to business development at her second CRO once her superiors saw that she could talk to people easily. She now works remotely from San Diego at Science Exchange, as a director of supplier relations.

"CROs are often fast-paced, cut-throat environments that teach you a lot in a short time."

A SHIFT IN FOCUS

Historically, CROs created positions that mirrored the services needed by the pharmaceutical sector. If a drug maker needed people to write journal manuscripts, for example, the CRO would supply medical writers — or project managers, or clinical-trial managers, or whatever a potential or existing client company might have required.

Now, however, large CROs are aiming to get ahead of the curve by providing data-management and data-analysis services, Getz says (see go.nature.com/2pnx2y5 and go.nature.com/2gg7sv9). Indeed, the landscape looks good for those who are skilled in these areas. CROs and biopharma plan to hire 25% more internal staff worldwide between now and 2020 for collecting, storing and making sense of the boatloads of data lurking in electronic health records (EHRs), social media and digital devices, according to a 2017 survey conducted by Tufts University's Center for the Study of Drug Development, where Getz is based.

Access to more-nuanced patient data is also enabling clinical-trial designs of greater complexity. As trial sponsors shift towards schemes that require different statistical and data-capture methods, researchers with those skills will be attractive to CROs, predicts Michael Winlo, chief executive of Linear Clinical Research, a mid-sized CRO in Nedlands, Western Australia, that specializes in early-stage clinical trials.

"Can you think about ways of accelerating a trial? If we drop from five to four patients, do we still get the statistical power we need?" Winlo says. "There's a lot of mental brainpower in trial design — in the statistics, in the writing, in making sense of the data."

Today, Parexel and other large CROs are hiring more data scientists — informaticians, epidemiologists and other people who can work with large data sets and extract insights from them. Candidates who can glean key information from insurance claims, EHRs and other real-world data will be in great demand, Schultz says. "Ten years ago, we had 2 people who could do this. Now we need maybe 100," he says. "It's become a skill set we actively seek."

Last year, Parexel bought Anolinx, a small, speciality CRO in Salt Lake City, Utah. Normally, Schultz adds, the target CRO would have been too small to acquire, but its data scientists had exactly the skills that Parexel sought.

Another example of innovation in trial outsourcing is Science 37, a company in Los Angeles, California, that functions as both a research 'site' and CRO, says co-founder

Belinda Tan. The company conducts virtual clinical trials through a telemedicine platform that allows researchers to easily find participants, who are able to avoid a trip to the clinic and get instructions from study staff through video calls at home. The platform serves as a data repository for all Science 37's trials, and staff members have access to some, depending on their role. Tan says that, for her as a physician, the repository acts like a clinical-trial EHR for participants.

Trial participants use mobile apps on smartphones provided by Science 37 to get their daily task list — for example, to complete a questionnaire or wait for a nurse to visit. To make these virtual trials work, Science 37 seeks not only conventional CRO candidates who have experience with clinical data, but also marketing and media specialists, web engineers, product designers, graphic designers and others.

Salaries for CRO employees vary widely depending on the level of education and job responsibilities. Clinical-research associates, who typically do not have PhDs, earn $50,000–65,000 on average in the United States, and clinical-research managers and clinical-research directors, who might have a doctorate, can earn more than $100,000.

Because they work with multiple clients, CROs tend to offer job stability — if one project fails or ends suddenly, the company can shift flexibly to another project with a different sponsor. And because the work is fast-paced and varied, employees can often broaden their skill sets and climb the career ladder more quickly than they would working at a pharmaceutical company.

Chakrabarti concedes that she misses one element of academic research. "You can follow a drug there from birth to clinic," she says. Conversely, CRO scientists often work with many different drug candidates at varying stages of development. "You have confidentiality agreements with these client companies, so you don't know anything about the compound. You do the assay but you don't know what happens later," she says. "Even when a molecule leaves a powerful impression — like, 'this is the strongest inhibitor I've ever seen' — your interest in a particular project has to stop with a particular deadline. This is what I find sad."

At least once so far, however, a chance run-in has brought the process full circle for her. At a cancer-therapeutics conference in Philadelphia, Pennsylvania, last October, Chakrabarti saw one of her clients presenting data about a familiar compound. She asked him if it was one that she had screened. Indeed, it was, he said, and the compound was heading into clinical trials. ∎

Esther Landhuis *is a freelance science journalist in the San Francisco Bay Area, California.*

COLUMN
Convert weaknesses into assets

Work out what you really enjoy doing, and pitch your skills accordingly, says **Lia Paola Zambetti**.

"I'm afraid I won't renew your contract. I am giving you as much advance notice as I can so that you can find something else."

Hearing these words from my supervisor's mouth left me reeling. As a native of Italy, and as a postdoctoral researcher in a nation outside the European Union, I had a visa that depended on my having a work contract. Without a job, I would have to leave the country shortly after the end of my contract.

Furthermore, the words felt like a death knell for my research career. Surely no one would ever hire me for a second postdoc when this one had failed to yield any research papers. What would I do in a few months' time when my postdoc ended? I was literally dizzy — I needed a strategy to find another position, and fast.

That was a tough week, but I am now grateful for that shocking announcement: it gave me clarity and enough time to make a plan.

The deadline made me think hard about my next steps. Somehow, I was able to start spelling out to myself what I emphatically did not want to do any more. It doesn't sound like the most logical step ever — surely, planning what you actually want to do makes more sense — but it was spectacularly helpful in clarifying my thoughts. Soon, I came up with a two-pronged strategy: first, look only for a research project that perfectly matches my wishes and skills; second, explore non-academic options as a real possibility — for the first time.

Because it looked increasingly likely that my future career was going to be outside academic research, I set out to turn my weaknesses into strengths. All the points that my supervisors and potential employers had highlighted as faults for a researcher — a poor publication record; no specific research niche; a tendency to 'waste time' reading papers from very different fields; and indulging my passion for writing — I aimed to turn into strengths for non-lab-based jobs.

Because I couldn't count on papers to speak for my research, I decided to network more. I converted my lack of a speciality into a 'broad and diverse background' and an ability to speak knowledgeably to scientists from different fields. My keen interest in writing, seen by some as a time sink, nudged me towards jobs in editing and science writing — something I had considered only as a vague dream.

I was not sure whether a good occupational fit for me existed, but I still had a few months to find out, so I set up informational chats with nearly everyone I could think of. And, for the first time ever, I was always straightforward about what I was — and was not — looking for in my new role.

One serendipitous talk on a Saturday led to a meeting with the director of the institute where I was doing my postdoc, which in turn led to an informal chat with a senior representative from the institute's marketing and corporate communications unit. She had been tasked with forming a science-communication team on an institution-wide level, and wanted to recruit a scientist.

Three months and two interviews later, the representative became my boss, and I had found my perfect fit in a role that focused on science communication and editing and that was completely away from the bench. As it happened, I also received an offer for a postdoctoral research project that aligned perfectly with my skills and interests. I regretfully felt obliged to decline it.

In the end, although it took all the time I had, I got not one but two great offers. And both matched my skills and interests — all because I had been clear about what I no longer wanted and because I had turned my weaknesses into strengths. ∎

Lia Paola Zambetti *is a senior project officer at the University of Sydney's Research Portfolio in Australia.*

GETTY

E-PLURIBUS

Time to play your part.

BY S. R. ALGERNON

---- **OFFICIAL SUMMONS** ----

Dear Citizen,

Pursuant to the Presidential Leadership Utilizing Representative Individual Brains in Unified Simulation (PLURIBUS) Act, you are hereby given notice of your selection to serve as this district's contribution to the collective leadership of the nation's Executive Branch. Your term of service shall last for **14 days** within the 3 months of **August** to **October** 2048.

You, along with 99 other registered voters, shall receive redacted briefings on domestic and foreign policy through a secure neural link. Your cognitive and affective responses to world events will inform the simulation's decision-making algorithms. Any thoughts, images or subvocalized speech during the recording period may be picked up by the link and incorporated into executive actions, communiqués to foreign powers and addresses to the nation. The people and their representatives turn to you for guidance and leadership.

If you do not already have neural-interface hardware installed, please report to an authorized implantation centre on or before **15 July 2048**. Alternatively, you may use the attached questionnaire to request a postponement or to be excused from service if:

1) You have already served two terms with PLURIBUS.
2) Your current residence is five or more light seconds from Earth.
3) You have a medical condition that contra-indicates neural-link implantation.
4) You are under 35 years of age and are not cognitively augmented to a mental age of 35 or greater.
5) You are over 70 years of age and are not taking senescence blockers.
6) You are currently serving a term with the legislative or judicial simulations.

As you prepare for the start of your term, please keep the following in mind:

● You will not know which 14 days will contribute to the PLURIBUS network. This has been shown to produce more natural decision-making and to reduce anxiety for PLURIBUS contributors.
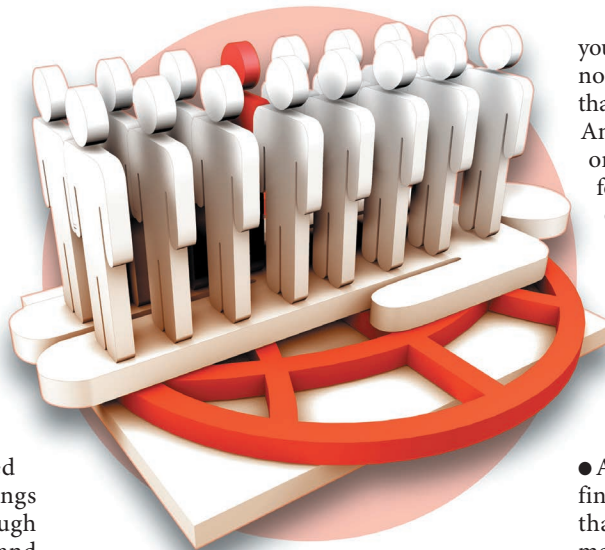
● Expect some cross-talk from other facets of PLURIBUS, especially in the first few days after implantation. Do not be alarmed. Most people quickly learn to differentiate their own thoughts from the stray thoughts of others.

● We request that you refrain from overly vigorous sexual activity during your term of office. We do our best to filter such things out, but we appreciate any effort you can make to ensure that PLURIBUS is not distracted or disturbed during what might turn out to be a crucial moment.

● Note that PLURIBUS comprises the neural patterns of 100 citizens at any one time. Please advise your friends and relatives not to hold you personally responsible for the nation's foreign or domestic policy.

● The Board of Elections recommends that you inform as few people as possible of your appointment until after your term expires, so that you are not subject to undue influence.

● With regard to the last point, please pay particular attention to conflicts of interest and receipt of gifts from foreign powers.

● During your term of office, it is likely that you will experience newfound historical knowledge, understanding of world events, and appreciation of economics and global trade. This is a normal effect of the neural link and is not a cause for alarm.

● Any sudden preference for tricornes, stovepipe hats or other eccentric headdress is not an effect of the neural link. We have no idea why this occurs.

● On any particular issue, you may find that PLURIBUS's final decision differs from your own view, or that your perspective was not adequately considered. Rest assured that PLURIBUS uses the latest algorithms. Any reports you may have read about bias or hacking are speculative and without foundation.

● During your term in office, you may feel overwhelmed by the pressures of the job. Remember that the burden does not rest entirely on your shoulders. You and your 99 other colleagues will depend on each other. It is not like the old days, when we entrusted the leadership of the executive branch to a single, flawed, human being.

● At the end of your term in office, you may find yourself feeling disconnected, or feel that your life no longer has the scope or meaning that it once had. Remember that life goes on after the presidency. Please consult the links at the end of this summons for advice on writing a presidential memoir and for links to several e-book providers that can help you create a personalized presidential library.

● Rarely, some PLURIBUS contributors report feeling that they are in the 'wrong body' or that they were 'switched' upon disengaging from the PLURIBUS network. Regrettably, we can only offer counselling to help you adjust to your new life after your term ends. As you may know, the Supreme Court is deadlocked on this issue. For the time being, we can only legally recognize 'you' on the basis of your corporeal form.

● You are free to use any social media accounts you may possess during your term in office. However, only the PLURIBUS collective has access to the official presidential accounts. Your individual e-mails, blog posts and tweets will not be considered presidential communications, for historical reasons.

As we would with any holder of this high office, we urge you to take the matter seriously and to set aside your own personal interests for the good of the country. You were selected on the basis of a profile derived from millions of ballots. You reflect the collective will of the people. No one person can do it alone; but together, we can — with foresight, wisdom, teamwork and empathy — achieve greatness. ■

S. R. Algernon *studied fiction writing and biology, among other things, at the University of North Carolina at Chapel Hill. He currently lives in Singapore.*

↻ **NATURE.COM**
Follow Futures:
🐦 @NatureFutures
f go.nature.com/mtoodm

ILLUSTRATION BY JACEY

How team science extends your scope

Modern cancer research demands input from collaborators across a broad spectrum of disciplines — not all of them scientific.

BY NEIL SAVAGE

When Daniel Stover was doing his postdoctoral research in a cell-biology lab at Harvard Medical School in Boston, Massachusetts, he ran into a problem. He was studying a type of breast cancer, trying to work out whether genetic differences between one part of a tumour and another contributed to the cancer's resistance to chemotherapy. He had plenty to work with — genetic information from hundreds of tumour samples — but no idea how to handle it all.

"I had generated an immense amount of sequencing data and couldn't find anybody to analyse it," says Stover, now an oncologist at the Ohio State University's Comprehensive Cancer Center in Columbus. So, with the help of a bioinformatician in the same lab, he started studying computational biology,

which became the focus of his studies. "I found that I loved working with data," he says. All the papers he published as a postdoc ended up being based on informatics, and now his own lab, which he set up last September, focuses on clinical computational oncology.

Stover's lab aims to fill the space between the computer experts who develop data-handling algorithms and the clinicians who focus on patient care, treatment and clinical trials. "In between, there's a gap, and we try to fill that void and take these amazing algorithms and apply them in clinical settings," he says.

Stover says the collaboration changed the direction of his career, in part because it gave him new skills that he could apply in working with other researchers.

Cancer research has become highly multi-disciplinary. The field now includes not just clinicians and molecular biologists, but also computational biologists, statisticians, nanotechnology experts and chemical engineers. And that creates challenges for all those researchers. How do they work with people who have different areas of expertise, each with its own basic assumptions and specialized language?

Nancy Krunic, who works for Novartis Pharmaceuticals in Cambridge, Massachusetts, heads the company's Future Precision Medicine Diagnostics group, which is developing assays, software and other technology to aid in diagnosis. "No one person or one department, or one lab, is going to have all the tools they need to tackle the problem," she says. "You absolutely

MAKING IT WORK
'Prenuptial agreements' for scientists

To help researchers to collaborate in multidisciplinary groups and anticipate difficulties in a project, the Office of the Ombudsman at the US National Institutes of Health came up with what it calls a prenuptial agreement for research teams. This lays out areas in which teams should reach agreements before problems arise (see go.nature.com/prenups). Here are some of the questions it suggests asking and answering before a collaboration begins in earnest:

- What are the scientific goals and expected outcomes of the project?
- When will the project be over?
- Who will write the reports?
- How will you decide what to do if discoveries made during the project change the direction of your research?
- Who will do the hiring, firing and supervising?
- How will credit and authorship be assigned?
- How will you make decisions about new collaborations or spin-off projects?
- What will you do about patents and intellectual property?
- Who will manage the data?
- What will happen if a collaborator changes job during the project?

need diverse backgrounds and subject-matter expertise."

Whether they're big pharmaceutical organizations or medical-device companies (Krunic previously worked at Luminex Molecular Diagnostics in Toronto, Canada), industry groups targeting cancer must form multidisciplinary teams, Krunic says, if they are to define and tackle problems in ways that are scientifically, clinically and commercially viable. As well as scientists and technologists, these teams will include people with expertise in, for example, marketing and regulatory issues, says Krunic.

BEYOND BIOLOGY
Programmes exist to promote cross-fertilization between disciplines. The US National Cancer Institute (NCI), for example, established a Physical Sciences in Oncology initiative in 2009 to team cancer biologists with physicists, mathematicians, chemists and engineers. Those disciplines come at cancer in a variety of ways. Chemical engineers devise new diagnostics and develop nanoparticles to carry drugs to tumours, or to act as contrast agents that make smaller tumours visible in imaging. Physicists and bioengineers study the effect of mechanical forces on tumour growth and behaviour, and mathematicians develop computational models to explain the complex interplay between different cancer cells, blood vessels, healthy tissue and drugs.

For example, researchers are working to understand the physical effects of a tumour's environment. How does an increase in tumour stiffness affect the shape and behaviour of the cells within it? And when a metastasizing cell deforms to squeeze through tight spaces, what does the increased pressure do to the cell's nucleus — does it, for instance, trigger processes that damage DNA? "It's not just the physical forces, but that's an important aspect of what's being studied," says Nastaran Zahir, director of the Physical Sciences in Oncology programme. Other projects include applying

mathematical approaches such as game theory to determine dosing strategies that will minimize the development of drug resistance, instead of applying the standard 'maximum tolerated dose' approach.

Zahir has experience of crossing disciplines. She earned her bachelor's degree in nuclear engineering, and studied plasma physics before moving into radiation biology and getting her PhD in bioengineering in a cancer research lab. So she's aware of the difficulties. "Biology has its own culture. Physics has a different culture," she says. "In physics, what you search for is sort of the ultimate truth — is there a law? But biology's very messy, and you don't necessarily have an exact process." Because biological processes change in response to new stresses, it's difficult to come up with laws for how a targeted cell would react to a cancer drug, for example.

LANGUAGE BARRIERS
To help bridge such gaps between disciplines, the NCI created the Science of Team Science programme. Kara Hall, a behavioural scientist who directs the initiative, says it's important for people to share knowledge with those from other disciplines in a comprehensible way. "That entails reducing the jargon that's being used, or finding ways to define that jargon as you go along," she says. It is often helpful to use analogies to explain key concepts in a field. It's also useful for researchers to engage in 'team learning', in which individuals are tasked with gaining in-depth information on a topic and bringing it back to their colleagues. Hall says teams should reflect on how well they function, by discussing, for instance, whether their meetings are sufficiently frequent and informative.

Hall says that people must be open when they approach specialists in other fields. It's important, she advises, to practise 'disciplinary humility' — to realize that all disciplines have both strengths and weaknesses, and to be willing to learn from fields other than your own.

Finding a safe common ground to ask questions can be difficult. "If I'm a psychologist collaborating with a geneticist, I may be afraid to ask a 'genetics 101' question because I might be seen as intellectually inferior," Hall says.

Other challenges in team science include the need for extra planning and management time, compared with individualized research. The approach can also require more team meetings and more travel, when collaborators are located across campus or at other institutions. And some institutional structures have not yet caught up with the concept, Hall says. Promotion and tenure committees tend to look mainly at the first and last authors of papers, she says. And that means they might not recognize how much a middle author has contributed — even though, in teams, middle authors play a crucial part in the research. Yet, Hall says, her programme's surveys found that trainees who had worked in multidisciplinary teams reported that their experiences had made them more competitive in the job market.

Defining the goals of a project, planning its implementation and working out in advance how to resolve conflict are all important parts of setting up a collaboration. The NCI offers the Team Science Toolkit, an online resource whereby researchers can share information and post news about funding opportunities and job openings (see go.nature.com/tstoolkit). It also helps to plan and support an annual Science of Team Science conference, which focuses on ways to make team-based research more effective. The next meeting runs from 21 to 24 May at the University of Texas Medical Branch in Galveston. And the US National Institutes of Health offers what it calls a "prenuptial agreement" to help scientists prepare for problems that can arise during a collaboration (see 'Making it work').

One early-career researcher taking a multidisciplinary approach is Viktor Adalsteinsson, who leads the blood-biopsy team in the cancer programme at the Broad Institute of MIT and Harvard in Cambridge. Although he earned his doctorate in chemical engineering, in 2015, Adalsteinsson knew from a young age that he wanted to help cure cancer. He did his PhD work at the Koch Institute for Integrative Cancer Research, which was set up at the Massachusetts Institute of Technology to bring biologists and engineers together under one roof. He helped to develop a system to isolate and sequence circulating tumour cells from blood samples, using his chemical-engineering education to deal with issues such as fluid dynamics and the amount of shear stress that cells could handle. Now in his own lab, he's trying to capture cell-free cancer DNA from blood to perform sequencing for precision medicine, reducing the need for invasive biopsies.

One of the ways in which Adalsteinsson and the people he works with stay up to date is through frequent meetings and seminars, at which various specialists talk about their work. Having a network of colleagues who can

explain research from other disciplines, or tell him whether a journal article is significant, is helpful, he says. "It's impossible to be an expert in every possible area, and knowing when to turn to others is really important."

POOL EXPERTISE

Sometimes the trick lies in knowing what not to read. "Being able to scan and reject a bunch of stuff is really important," says Heather Parsons, a medical oncologist and physician at the Dana–Farber Cancer Institute in Boston, who specializes in breast cancer and its biomarkers. She, too, emphasizes the importance of having a network of experts, developed through university and work, with whom you can discuss questions.

Parsons collaborates with Stover and Adalsteinsson on the liquid-biopsy work. "I like very much being part of this kind of a team," she says, "but it requires that you don't have an enormous ego and you don't mind asking about things you don't understand."

At Stanford University School of Medicine in California, Guillem Pratx gets members of his physical oncology lab to take part in a journal club. They meet for an hour or so to focus on a particular paper, allowing people from different disciplines to gain a good understanding of its importance. He also requires them to attend meetings outside their field to broaden their knowledge. With enough exposure, he says, scientists can become comfortable with the terminology and concepts used in other areas. "I notice the more I sit in these talks, the more I understand," Pratx says. "It's like learning a new language."

Pratx did his undergraduate and graduate studies in electrical engineering, and during his PhD studies he worked in a radiology lab, using graphics techniques from computer games to improve the processing of medical images. He did his postdoctoral research in radiation oncology, and he feels that using postdoc time to learn about an area outside one's core speciality can pay off. It can be difficult to be hired by a lab that specializes in a field far removed from yours, he acknowledges. But if there's some overlap, it can add valuable expertise.

Pratx's lab, which includes scientists with backgrounds in physics, engineering, chemistry and biology, develops instruments, probes and algorithms for cancer imaging. The team is studying how the luminescence generated when therapeutic radiation hits tissue can be used to carefully aim the otherwise-invisible beam. One challenge for such multidisciplinary teams is communicating to different members how they can tackle a problem, he says. Biologists often struggle to understand what questions mathematical models can ask concerning the large data sets generated in cancer research — sets that include not only genomic and proteomic sequences, but also imaging results and environmental information from medical records. It's important that there's someone in the group who understands which statistical methods are best applied to particular types of data, and what the results do and don't show, says Pratx.

On the flip side, he thinks that engineers can focus too much on trying to come up with innovative techniques, and are sometimes less interested in applying what others have already developed. It's not enough for insights gleaned from data to be new, he says. They also have to be biologically relevant.

One problem that Pratx sees is the one Stover experienced. Although the growth in data is increasing the need for computational specialists in cancer research, the competition from other fields for people with those skills is strong.

MATCHMAKERS

Early-career researchers interested in forming collaborations need to network with people from other fields, and one obvious way is to attend conferences in those fields. But Jennifer Podesta, a molecular biologist and a specialist in the use of nanotechnology for drug delivery, says that simply attending a conference isn't enough. "Do a little bit of homework, and go in very much with an agenda of 'who it is I want to meet and what do I want to get out of it?'" she says. "It's remarkable how many people think they can show up, scrunch over and stand in the corner, and come away from it complaining that they didn't meet a collaborator."

Podesta, who runs the Cancer Research UK Centre at Imperial College London, recommends working out the type of scientist you need for the project you have in mind, and then approaching department heads in your own university to see who they think might fit. Funding managers also tend to have a broad knowledge of which researchers have what expertise, and are usually happy to play matchmaker.

Getting funding for cross-disciplinary projects can be challenging, especially for someone who hasn't yet established a reputation, so Podesta suggests looking for small sums of money internally, to fund a pilot project with a new collaborator. Such projects demonstrate that members of the team can work together and produce viable ideas, making them more attractive to funding agencies. The NCI's Physical Sciences in Oncology programme provides funding specifically for pilot projects.

Trying to keep up with a field as dynamic as cancer research is daunting. "We have so much information within our reach, and new discoveries are being made every day," Adalsteinsson says. The key to tackling all that information, Pratx says, is to overcome the tendency of many scientists to think they need to learn everything themselves. "I think it's an important skill when you're able to say, 'Maybe I don't need to be an expert in computer modelling. I can maybe work with somebody else,'" he says. ∎

Neil Savage *is a freelance science and technology writer in Lowell, Massachusetts.*